DEALING WITH LINGUISTIC MISMATCHES FOR AUTOMATIC SPEECH RECOGNITION

BY

XUESONG YANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Mark Hasegawa-Johnson, Chair
Professor Thomas S. Huang
Associate Professor Paris Smaragdis
Associate Professor Chilin Shih

ABSTRACT

Recent breakthroughs in automatic speech recognition (ASR) have resulted in a word error rate (WER) on par with human transcribers on the English Switchboard benchmark. However, dealing with linguistic mismatches between the training and testing data is still a significant challenge that remains unsolved. Under the monolingual environment, it is well-known that the performance of ASR systems degrades significantly when presented with the speech from speakers with different accents, dialects, and speaking styles than those encountered during system training. Under the multi-lingual environment, ASR systems trained on a source language achieve even worse performance when tested on another target language because of mismatches in terms of the number of phonemes, lexical ambiguity, and power of phonotactic constraints provided by phone-level n-grams.

In order to address the issues of linguistic mismatches for current ASR systems, my dissertation investigates both knowledge-gnostic and knowledge-agnostic solutions. In the first part, classic theories relevant to acoustics and articulatory phonetics that present capability of being transferred across a dialect continuum from local dialects to another standardized language are re-visited. Experiments demonstrate the potentials that acoustic correlates in the vicinity of landmarks could help to build a bridge for dealing with mismatches across difference local or global varieties in a dialect continuum. In the second part, we design an end-to-end acoustic modeling approach based on connectionist temporal classification loss and propose to link the training of acoustics and accent altogether in a manner similar to the learning process in human speech perception. This joint model not only performed well on ASR with multiple accents but also boosted accuracies of accent identification task in comparison to separately-trained models.

To my parents, my wife, and my little one,
for their love and support.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor Prof. Mark Hasegawa-Johnson for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides, I would like to thank the rest of my multidisciplinary advisory committee: Prof. Thomas Huang and Paris Smaragdis from ECE, and Prof. Chilin Shih from Linguistics, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

My sincere thanks also goes to my awesome industry mentors and collaborators: Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, and Bhuvana Ramabhadran from IBM Research; Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng from Microsoft Research; Gregg Wilensky, Trung Bui, Hailin Jin, and Walter Chang from Adobe Research; Anastassia Loukina and Keelan Evanini from ETS Research. Without their precious support, it would not be possible to further explore the depth and breadth of product-driven speech and language techniques.

I thank my fellow IFP labmates: Kaizhi Qian, Yang Zhang, Shiyu Chang, Amit Das, Di He, Xiang Kong, Sujeeth Bharadwaj, and Po-Sen Huang, for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last six years. Also, I thank visiting professors in our lab: Yanlu Xie and Zhijian Ou, for enlightening me the first glance of research, and for the encouragement of continuing my Ph.D. study.

Last but not the least, I would like to thank my family: my parents, my wife, and my son, for supporting me spiritually throughout researching new ideas, writing this thesis, and taking adventures in this American life in general.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Automatic speech recognition (ASR), as the system transcribing human speeches into texts, has been an active research area for decades. It becomes an essential bridge for better communications between human and machines. In the early stage of ASR development, people prefer to use keyboards and mice as their main input methods rather than to use voice interfaces in that ASR performance was not reliable. Recent research progress has resulted in a much lower word error rate (WER) ever since the interest in hybrid ASR systems with deep neural networks (DNNs) and hidden Markov models (HMMs) resurged [1]. Breakthroughs on English Switchboard benchmark are achieved on par with human transcribers [2,3]. Figure 1.1 illustrates the evolution of ASR performance in terms of WER. Speech-driven products, such as Amazon Alexa, Google Home, Microsoft Cortana, and Apple Siri, grow rapidly in the market, but still, they are not broadly accepted in our daily lives due to the deficiency of front-end ASR systems.

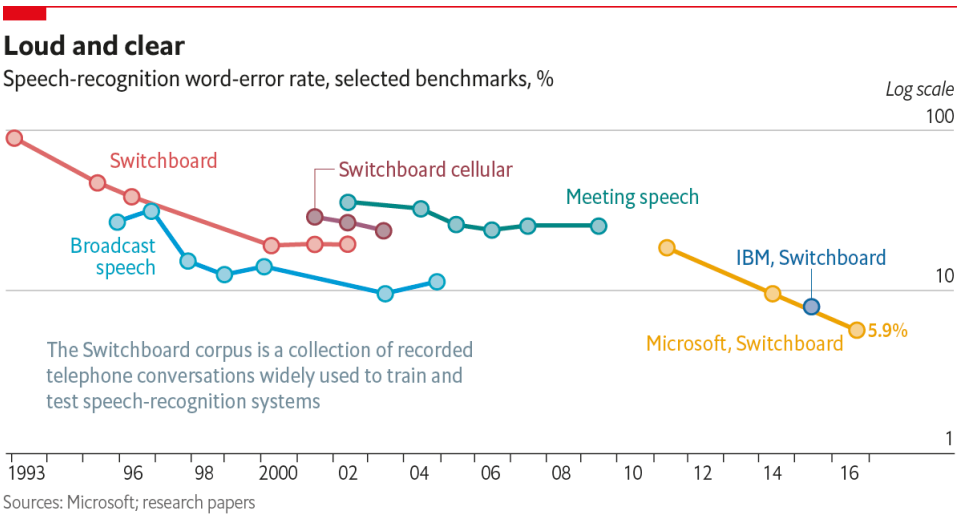


Figure 1.1: Evolution of ASR performance on selected benchmarks[1].

[1] https://www.economist.com/technology-quarterly/2017-05-01/language

The state of the arts of ASR systems remain to highly rely on the HMM paradigm in a hybrid of expressive statistical models, such as Gaussian mixture models (GMMs), resurgent DNNs, and various recurrent neural networks (RNNs). From the perspective of statistical learning theory, both training and testing examples are assumed to be drawn from the same feature space and the same probability distribution, otherwise, those statistical models would be adversely affected by the mismatches [4, 5]. In practice, however, this standard assumption often does not hold—the training and testing data distributions may somewhat differ. In the context of ASR applications, the mismatches between training and testing examples mainly result from different conditions in environmental noise, transducers, channels, speaker characteristics, and language dialects. This dissertation drills down to explore feasible solutions to the challenging problems of linguistic mismatches for ASR systems.

1.1 Linguistic Mismatches: Interpretation from Dialect Continua

Linguistic mismatches are variations of a language that the speech community accepts and typically do not lead to human communication problems. However, such mismatches often lead to ASR errors. One of the major types of linguistic mismatches that pose problems for ASR is the subtle dialect variations human listeners can adapt to quickly. Very often, such dialectal variations form a dialect continuum where the dialectal distance is reflected in the geological distance where close-by geographical areas have similar language varieties that differ slightly but areas further apart show more diverse language variations. As the geological and linguistic distance[2] grow, the dialects may no longer be mutually intelligible. Naturally, many factors affect the forming of a dialect continuum and the linguistic distance may not reflect geological distance due to political, economic, transportation and migration patterns throughout the history. When a dialect continuum is observable, dialectologists draw lines to separate areas that differ with respect to some features. Standard varieties in each area together with its dependent varieties are considered as a "language", while those dependent varieties are considered

[2]Linguistic distance is a measure of how different one language or dialect is from another based on mutual intelligibility, i.e. the ability of speakers of one language to understand the other. The higher linguistic distance, the lower is the level of mutual intelligibility [6].

as “dialects”. Figure 1.2 illustrates the concept of a dialect continuum. Speakers of local varieties usually read and write in a related standardized form, and use it for official purposes, for example, on radio or television; the standardized form may change in their pronunciation realizations due to another independent culture status [7]. Standard Dutch and standard German are exemplars of standardized forms in the West Germanic language family in that they are not closely linked with regard to their ancestral dialects and hence they do not show a high degree of mutual intelligibility when spoken. Local dialects of such west Germanic continuum are oriented towards either Dutch or German, depending on which side of the state border they are spoken.

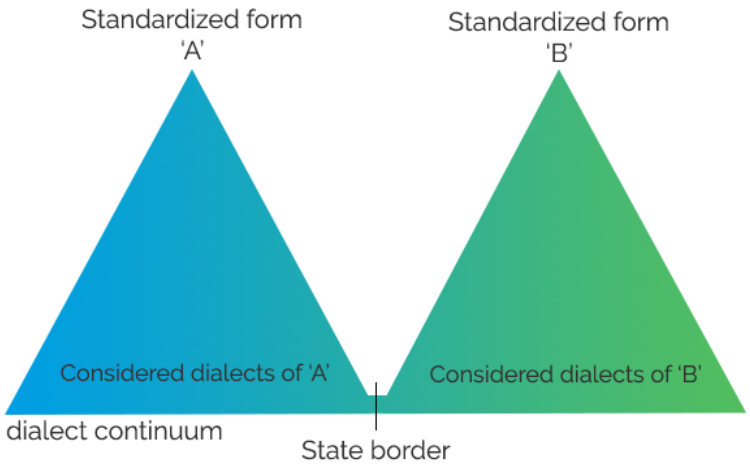


Figure 1.2: Illustration of a dialect continuum.

From a dialect continuum’s point of view, under the monolingual environment, it is well-known that the performance of ASR systems degrades significantly when presented with the speech from speakers with different accents, dialects, and speaking styles than those encountered during system training [8]. Under the multi-lingual environment where linguistic distance in the dialect continuum grows across the state border as shown in Figure 1.2, ASR systems trained on a source language achieve even worse performance when tested on another target language because of mismatches in terms of number of phonemes, lexical ambiguity, and power of phonotactic constraints provided by phone-level n-grams [9]. In contrast to human speech recognition, human listeners usually well perceive the voices even with linguistic mismatches and they are capable of outperforming machines by a large margin. Figure 1.3 demonstrates the above observations of differences on the accented speech corpora Librispeech [10] by comparing human speech recognition

and the *DeepSpeech2* system [11] that is one of state-of-the-art end-to-end ASR systems. That being said, it is still a long row to hoe for ASR techniques before approaching the ability of human speech recognition.

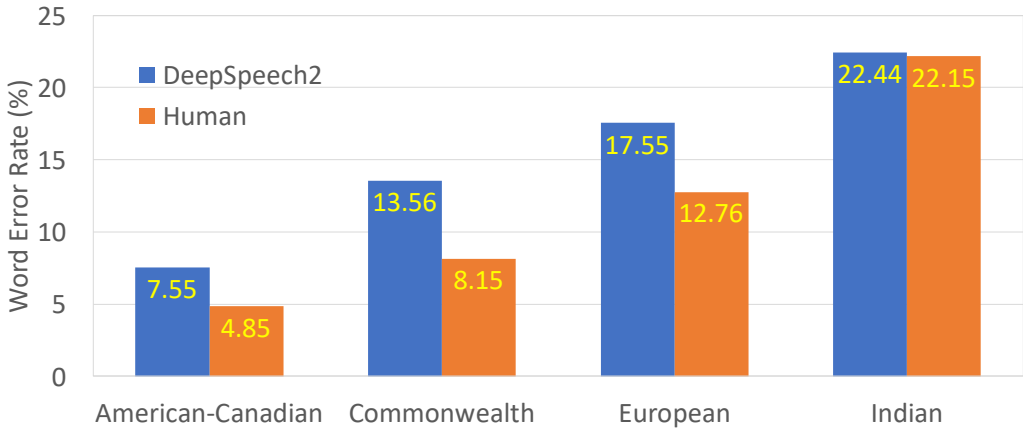


Figure 1.3: Comparison of human speech recognition and *DeepSpeech2* on the accented English corpora Librispeech. *European* category represents countries in Europe where English is not a first language. *Commonwealth* represents the accents from British, Irish, South African, Australian and New Zealand.

1.2 Contribution

Inspired by the interpretation of linguistic mismatches from the perspective of dialect continua, this dissertation explored *knowledge-gnostic* solutions to the issues of linguistic mismatches for current ASR systems by leveraging prior knowledge of acoustic, articulatory, and auditory phonetics that presents capability of being transferred across a dialect continuum from local dialects to another standardized language. This dissertation also investigated *knowledge-agnostic* solutions in an end-to-end (E2E) fashion, especially by leveraging benefits from connectionist temporal classification (CTC) without any needs of prior phonetics knowledge. All discoveries are summarized from my peer-reviewed papers [12–21] along with their supplementary technical reports, and would be elaborated in the following.

Acoustic Landmarks Bridge Mismatches Across Languages

Acoustic landmarks exploit quantal nonlinear articulator-acoustic relationships, which identify times when acoustic patterns of linguistically motivated distinctive

features are most salient. Acoustic cues extracted in the vicinity of landmarks may, therefore, be more informative for identification of articulator manner changes and classification of distinctive features than other cues extracted from other times in the signal. We conducted ASR decoding experiments on TIMIT by a heuristic approach of weighting acoustic likelihood scores of speech frames. We found that speech frames locating at landmarks are more informative for recognition than other frames; our system can maintain the same phone error rate (PER) when even a half of non-landmark frames are dropped during decoding process [12, 13]. Acoustic landmark theory is claimed to be suitable for any dialects and languages. Dialects may differ in phonetic features, but landmarks are invariant to different dialects. Therefore ASR systems could adapt to dialect variations by extracting essential acoustic information at landmarks. We validated such portability of acoustic landmark theory to any languages in applications of pronunciation error detection [14].

Previous experiments and analysis are conducted in the sense that acoustic landmark positions are known in advance so that accurate information related to manners and places of articulation can be further extracted according to their acoustic correlates. We believe this assumption is actually not realistic since transcriptions with landmark annotations are rather scarce in any languages other than English[3]. Labeling accurate acoustic landmarks requires annotators with solid background knowledge of phonetics and phonology. It becomes an obstacle to build up knowledge base accessible to any languages, such that the potential benefits of acoustic landmarks are discarded. In order to fill gaps of insufficient landmark labels for any languages, we implemented deep neural models to detect manner changes of articulation in English. According to our previous findings [14] where acoustic landmark positions are demonstrated to be capable of being transferred cross-lingually, these landmark detectors trained on English is suitable to apply to completely new speech in completely new languages as well [15]. We also attempted to find an alternative approach to obtain landmark labels by analyzing the correlation between spiky CTC predictions and acoustic landmark positions based on peak detection algorithms [16]. The preliminary experiments suggested a promising way of landmark annotation.

[3]Only English TIMIT and its noise-perturbed NTIMIT provide detailed landmark information

Distinctive Feature Classifiers Are Transferrable Crosslingually

Differences can be observed in the acoustic manifestation of the same feature in different languages. The acoustic correlates of categorical distinctive features, especially of the articulator-bound features, are still most salient in the vicinity of landmarks. Building classifiers on these expressive acoustic correlates may contribute to distinguishing categorical distinctive features in any languages, which furthermore may help to improve the performance of ASR systems with the combination of acoustic signal observations and phonetic information in multi-lingual scenarios. We designed multiple classifiers anchored at phonetic landmarks to detect the consonantal voicing feature. Experiments demonstrated the advantage of these classifiers—they can be transferred cross-lingually while without loss of detection accuracy [17].

Joint Modeling of Acoustics and Dialects Using CTC

Local varieties of pronunciations in a standardized form typically share acoustic representations that are universal to any dialects while they can be distinguished by individual characteristics that preserve dialect-specific knowledge. Ideal configuration of an end-to-end model assumes to utilize all sources of dialects to train the model that captures shared representations across dialects. But this strategy doesn't work well since it ignores the mismatches across different domains. Another extreme configuration only needs dialect-specific data while missing common patterns shared across domains.

Previous approaches of dealing with accented speech do not explicitly utilize accent information during the training procedure of acoustic models, but do so only indirectly, for example, through different target phoneme sets for various accents. These solutions contrast sharply with the way in which humans memorize the phonological and phonetic forms of accented speech. We considered accent identification as an auxiliary task and proposed to link the CTC training of acoustic models and cross-entropy training of accent identification models altogether [18]. This auxiliary task helped by introducing extra accent-specific information, which augmented general acoustic features.

The CTC algorithm learns accurate acoustic models without time-aligned phonetic transcription, but sometimes it fails to converge, especially in resource-constrained scenarios. We tailored a new set of acoustic landmarks and leveraged

new target label sequences mixed with both phones and manner changes of articulation to help CTC training converge more rapidly and smoothly while also reducing recognition errors. We also investigated the sensitivity of our approaches to the size of training data on subsets of TIMIT (smaller corpora) and WSJ (a larger corpus). Consistent performance gain demonstrated that label augmentation with acoustic landmarks generalizes to larger and smaller training datasets, and we believe this is the first work that applies acoustic landmark theory to a mid-sized ASR corpus [19].

1.3 Organization

The rest of this dissertation will describe details of the above contributions. Chapter 2 introduces background knowledge related to acoustic and articulatory phonetics that serves as the foundation for the following chapters. Chapter 3 demonstrates the impact of speech frames located in the vicinity of acoustic landmarks by weighting acoustic likelihood scores of speech frames during the decoding process. Chapter 4, Chapter 5, and Chapter 6 discuss the portability of acoustic landmark theory to any languages and apply the advantages to the task of pronunciation error detection. Chapter 7 bridges linguistic mismatches by analyzing the effectiveness of articulatory distinctive feature classifiers across multiple languages. Chapter 8 explores CTC-based ASR systems on multiple accents by jointly modeling of acoustics and accents. Chapter 9 further explores the benefits of target label augmentation approaches with acoustic landmarks for rapid and stable CTC training. Chapter 10 reemphasizes my contributions towards the goal of dealing with linguistic mismatches for ASR and discusses future directions.

CHAPTER 2

BACKGROUND

2.1 The Quantal Nature of Speech

The human vocal tract is a generator of speech, and likewise, the human auditory system is a receiver of speech. We are interested in the nature of two kinds of relations regarding vocal tract, auditory system, and speech—relations between vocal tract configurations or states and the properties of speech generated from these articulations; relations between acoustic parameters of speech and auditory responses to the speech described by these parameters. Stevens discovered similar nonmonotonic transformations for these two kinds of relations: *acoustic-articulation* [22] and *auditory-acoustic* [23].

Acoustic-articulation can be further elaborated schematically in Figure 2.1 which shows the relationship between some acoustic parameter in speech generated from the vocal tract and some articulatory parameter controlled by speakers. Region II represents large changes in the acoustics for small shifts in articulation; Region I and Region III show two plateaus in the curve indicating that the acoustic parameter remains relatively stable when small modifications are made in the articulation. The difference between values of acoustic parameters in Region I and Region III is large, which indicates that a significant acoustic contrast or a rather abrupt change between these two regions occurs in their intermediate Region II.

Auditory-acoustic applies the same non-monotonic curve as *acoustic-articulation* described in Figure 2.1. In this case, the dependent variable (y-axis) is some parameter of the auditory response measured by psychophysical procedures or electrophysiological methods, and the independent variable (x-axis) is some acoustic parameter that is controlled by a speaker. Region II here is considered as a transition band showing that the auditory response shifts from one type of pattern to another as the acoustic parameter changes across this region.

In summary, some articulatory states or gestures give rise to well-defined patterns

of auditory response from a human listener, so that these patterns are not very sensitive to small perturbations in the articulation while being distinctive in the sense that a significant change or qualitative shift occurs in the auditory response if some articulatory parameter passes through a threshold region. The relation of *acoustic-articulatory* or *auditory-acoustic* posits quantal attributes characterized by rapid changes in state over some threshold regions while keeping stability over other regions. Stevens suggested that the quantal relationship is a principal factor shaping the inventory of articulatory gestures and their acoustic generation that are used to distinguish phonetic segments in language [24]. The relationship between articulatory and acoustic attributes captures the acoustic correlates of the distinctive features. A speech utterance can be represented as a sequence of multi-dimensional and categorical distinctive features. At the plateau-like regions, some changes in these features will not result in a significant modification of attributes of the sound pattern; when one of the distinctive features crosses through the threshold region, rapid and abrupt changes will occur in the relevant acoustic parameter. In this context, the time is defined as an *acoustic landmark* by Stevens et al [25] when an acoustic event of a rapid change occurs. Furui [26] and Ohde [27] have made almost the same observations of these relationships when studying Japanese syllables and children's speech.

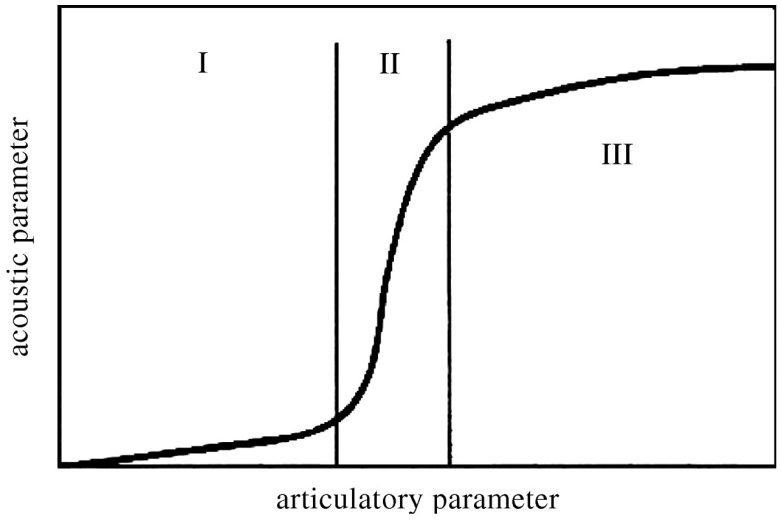


Figure 2.1: Quantal relationships between an articulatory parameter of the vocal tract and the acoustic output. Regions I and III are acoustically stable with respect to perturbations in the articulatory parameter, whereas Region II contains acoustically abrupt changes. This figure is originally from [24].

2.2 Distinctive Features

Distinctive features concisely describe sounds of a language at a subsegmental level, and they have direct relations to acoustics and articulation. These features are typically quantized as binary values that encode perceptual [28], phonological [29], and articulatory [23] speech sounds. A collection of these binary features can distinguish each phonetic segment from all others in a language. For example, the vibration behavior of the vocal cord determines if the speech sound is "voiced" ([+periodicGlottalSource]) or "unvoiced" ([-periodicGlottalSource]), and the velum helps to determine if it is a "nasal" ([+nasal]) or "non-nasal" ([-nasal]) sound.

Distinctive features in a language often have articulatory, acoustic, or perceptual correlates that are similar to those in other languages. They are sufficient to characterize all sounds in all languages so that phoneme systems can be derived for any languages in the world. Stevens [30, 31] suggested to form a set of approximately language-independent distinctive features: if an acoustic or articulatory feature is used to distinguish phonemes in at least one of the languages of the world, then that feature may be considered to define a language-independent distinctive feature. Each phoneme of a language is represented by a unique binary vector of language-independent distinctive features. ASR may distinguish two different allophones of the same phoneme as distinct phones. In most cases, the distinctions among phones can be coded using distinctive features borrowed from another language, or equivalently, from the language-independent set. Table 2.1 describes a collection of features[1] suitable for any languages.

The ASR community has explored a number of encodings strategies similar to distinctive features, such as *articulatory features* and *speech attributes*. *Articulatory features* can help to classify sounds of a language, but they take on mostly digital values (e.g. velum position) or continuous values (e.g. horizontal position of the dorsum). Generally speaking, distinctive features contains the quantized values of articulatory features. Many studies have focused on articulatory features mainly because it has superb advantages for dealing with impacts of noisy and reverberant environment [32–34], providing a compact representation of pronunciation variability [35], and compensating variabilities across multiple languages [36–38]. *Speech attributes*, on the other hand, are the superset of distinctive features. they are deliberately defined to introduce other purposes to speech recognition. For example, Lee et al [39] defined quite broad speech attributes to bridge the perfor-

[1] Distinctive features across languages is retrieved from `http://phoible.org`

mance gap between ASR and human speech recognition. Those attributes include a collection of information beyond distinctive features, such as acoustic cues signaling a speaker's gender, accent, emotional state, and other prosodic, meta-linguistic, and para-linguistic messages.

Table 2.1: Language-independent distinctive features.

advancedTongueRoot	anterior	approximant
back	click	consonantal
constrictedGlottis	continuant	coronal
delayedRelease	distributed	dorsal
epilaryngealSource	fortis	front
high	labial	labiodental
lateral	long	low
loweredLarynxImplosive	nasal	periodicGlottalSource
raisedLarynxEjective	retractedTongueRoot	round
short	sonorant	spreadGlottis
stress	strident	syllabic
tap	tense	trill

2.3 Acoustic Landmarks and Its Application on ASR

The name of "acoustic landmarks" was firstly introduced by Stevens et al [25] in 1992, and its theory originated from experimental results of human speech perception which demonstrated that perceptual sensitivity to acoustic events is not linearly correlated in either time or frequency domain. It exploits quantal nonlinearities in articulatory-acoustic [22] and auditory-acoustic [23] relations to define instances in time when abrupt changes occur in speech articulation, in the speech spectrum, or in a speech auditory response. Landmark theory states that human perception of phonemes corresponds to acoustic cues anchored temporally in the vicinity of landmarks where salient distinctive features can be detected. As opposed to modern statistical ASR where each frame is treated with equal importance, landmark theory proposes that there exist information rich regions in the speech utterance and that we should focus on these regions more carefully. These regions of interest are anchored at acoustic landmarks. Landmarks are instantaneous speech events where distinctive features are most clearly signaled.

Hasegawa-Johnson [40] defined a set of landmarks including consonant releases

and closures (at phone boundaries) and vowel/glide pivot landmarks (near the center of the corresponding phones). In contrast, Lulich [41] argued that the center of vowels and glides are not as informative and should not be considered as landmarks. He defined, instead, formant-subglottal resonance crossing, which is known to sit between boundaries of [-back] and [+back] vowels, to be more informative. Wang et al [42] showed that the latter proposal help to improve performance for automatic speaker normalization tasks. However, the former definition[2] of landmarks by Hasegawa-Johnson [40] is more suitable for ASR tasks and it provides a better approximation of the typical timing of the spectro-temporal events discovered in Liu's earlier work [43]. Later works [17, 44] achieved comparable performance by annotating these landmarks right on phone boundaries. Figure 2.2 illustrates an example of landmark annotations for the word "symposium" selected from an utterance on TIMIT[3].

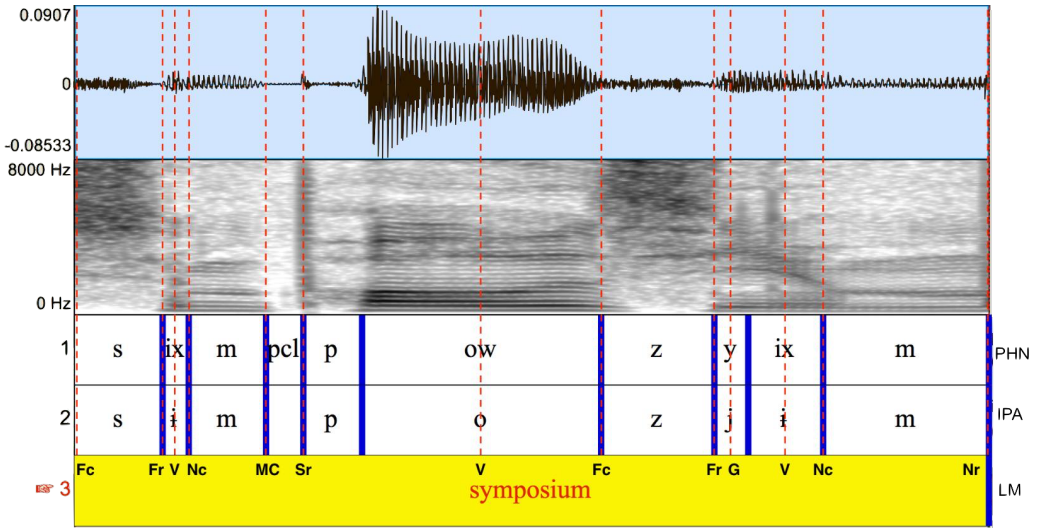


Figure 2.2: Acoustic landmark labels for the word "symposium". The audio is transcribed using TIMIT phone symbols (PHN) and international phonetic alphabet symbols (IPA). Landmark positions are labeled in red dashed lines where landmark types are detailed. Fc and Fr are the closure and release for fricatives; Sc and Sr are the closure and release for stops; Nc and Nr are the closure and release for nasals; V and G are the vowel pivot and glide pivot; MC is a manner change.

Many other works have focused on accurately detecting acoustic landmarks. Some of them assumed that landmarks correspond to the temporal extrema of speech energy or energy changes in particular frequency bands, such as consonantal

[2] A small number of the pivot and release landmarks are defined at salient locations where a time-delay by +33% or a time-advance by -20% of the phone duration exists.

[3] `TIMIT/TRAIN/DR1/FSMA0/SX361.WAV`

landmarks detection [43], vowel landmarks detection [45], consonant voicing classification [46], and places classification of articulations [47–50]. Support vector machines (SVMs) demonstrated superb performance on detecting stop release landmarks over hidden Markov models [51] so that SVMs are further explored to detect all other landmarks [52, 53]. Qian et al [54] continued to apply SVMs-based approaches for stop consonant detection by extracting more expressive acoustic cues using the local binary patterns resulting and achieved the accuracy above 95%. Xie et al [55] extended Niyogi's work [51] and discovered that distinctive feature classifiers are also beneficial to detect landmarks. Our previous paper [17] attempted to detect consonant voicing, one of distinctive features, by applying a binary classifier based on a convolutional neural network with MFCCs and additional energy-based acoustic features. This system was trained on the English corpus TIMIT [56], but tested on other languages such as Spanish and Turkish. We achieved around 85% accuracy for all languages.

Acoustic landmarks have been used in a variety of ASR system architectures. These systems, without considering the mechanism used for landmark detection, can be clustered into two types. The first type of system, as described in [43,57,58], computes a lexical transcription directly from a collection of detected distinctive features. Due to the complexity of building a fully decoding mechanism on distinctive features, some of these systems only output isolated words. However, other systems (e.g., work from [57]) have full HMM back-ends that can output word sequences. The other type of system, as described in [44], conducts landmark-based re-scoring on the lattices generated by MFCC-based hidden Markov models. Acoustic likelihoods from the classic ASR systems are adjusted by the output of the distinctive feature classifiers. Many landmark-based ASR systems demonstrated the word error rates slightly [44] or even significantly [32] better than the statistical ASR models, especially in noisy conditions.

CHAPTER 3

ACOUSTIC LANDMARKS CONTAIN RICH INFORMATION ABOUT PHONE STRINGS

Most mainstream ASR systems consider all acoustic feature frames equally important. However, acoustic landmark theory is based on a contradictory idea, that some frames are more important than others. Acoustic landmark theory exploits quantal non-linearities in the articulatory-acoustic and acoustic-auditory relations to define landmark times at which the speech spectrum abruptly changes or reaches an extremum; frames overlapping landmarks have been demonstrated to be sufficient for speech perception. In this chapter, we conduct experiments on the TIMIT corpus, with both Gaussian Mixture Model (GMM) and Deep Neural Network (DNN) based ASR systems and find that frames containing landmarks are more informative for ASR than others. We find that altering the level of emphasis on landmarks by re-weighting acoustic likelihood tends to reduce the phone error rate (PER). Furthermore, by leveraging the landmark as a heuristic, one of our hybrid DNN frame-dropping strategies maintained a PER within 0.44% of optimal when scoring less than half (45.8% to be precise) of the frames. This hybrid strategy out-performs other non-heuristic-based methods and demonstrate the potential of landmarks for reducing computation.

3.1 Introduction

Ideas from speech science—which may have the potential to further improve modern ASR techniques—are not often applied to them [2]. Speech science has demonstrated that perceptual sensitivity to acoustic events is not uniform in either time or frequency. Most modern ASR uses a non-uniform frequency scale based on perceptual models such as critical band theory [59]. In the time domain, however, most ASR systems use a uniform or frame synchronous time scale: systems extract and analyze feature vectors at regular time intervals, thereby implementing a model according to which the content of every frame is equally important.

Acoustic landmark theory [23, 60] is a model of experimental results from speech science. It exploits quantal nonlinearities in articulatory-acoustic and acoustic-auditory relations to define instances in time , also knonw as landmarks, where abrupt changes or local extrema occur in speech articulation, the speech spectrum, or a speech perceptual response. Landmark theory proposes that humans perceive phonemes in response to acoustic cues, which are anchored temporally at landmarks, i.e., that a spectrotemporal pattern is perceived as the cue for a distinctive feature only if it occurs with a particular timing relative to a particular type of landmark. Altering distinctive features alters the phone string; distinctive features in turn get signaled by different sets of cues anchored at landmarks.

The theory of acoustic landmarks has inspired a large number of ASR systems. Acoustic landmarks have been modeled explicitly in ASR systems such as those reported by [44,57,58]. Many of these systems have accuracies comparable to other contemporaneous systems—in some cases, even returning better performance [44]. However, published landmark-based ASR with accuracy comparable to the state of the art has higher computation than the state of the art; conversely, landmark-based systems with lower computational complexity tend to also have accuracy lower than the state of the art. No implementation of acoustic landmarks has yet been demonstrated to achieve accuracy equal to the state of the art at significantly reduced computational complexity. If acoustic landmarks contain more information about the phone string than other frames, however, then it should be possible to significantly reduce computational complexity of a state of the art ASR without significantly reducing accuracy, or conversely, to increase accuracy without increasing computation, by forcing the ASR to extract more information from frames containing landmarks than from other frames.

We assume that a well-trained frame-synchronous statistical acoustic model (AM), having been trained to represent the association between Mel-frequency cepstral coefficients (MFCC) features and triphones, has also learned sufficient cues and necessary contexts to associate MFCCs and distinctive features. However, because the AM is frame-synchronous, it must integrate information from both informative and uninformative frames, even if the uninformative frames provide no gain in accuracy. The experiments described in this paper explore whether, if we treat frames containing acoustic landmarks as more important than other frames, we can get better accuracy or lower computation. In this work, we present two methods to quantify the information content of acoustic landmarks in an ASR feature string. In both cases, we use human annotated phone boundaries to label the location of

landmarks. The first method seeks to improve ASR accuracy by over-weighting the AM likelihood scores of frames containing phonetic landmarks. By over-weight, we mean multiplying log-likelihoods with a value larger than 1 (Section 3.2.1). The second method seeks to reduce computation, without sacrificing accuracy, by removing frames from the ASR input. Removing frames makes the computational load decrease, but usually causes accuracy to decrease also; which frames can be removed that cause the accuracy to drop the least? We searched for a strategy that removes as many frames as possible while attempting to keep the Phone Error Rate (PER) low. We show that if we know the locations of acoustic landmarks, and if we retain these frames while dropping others, it is possible to reduce computation for ASR systems with a very small error increment penalty. This method for testing the information content of acoustic landmarks is based on past works [61–63] that demonstrated significantly reduced computation by dropping acoustic frames, with small increases in PER depending on the strategy used to drop frames. In this paper we adopt the PER increment as an indirect measure of the phonetic information content of the dropped frames.

If the computational complexity of ASR can be reduced without sacrificing accuracy, or if the accuracy can be increased without increasing the computational load, these findings should have practical applications. It is worth emphasizing that this work only intends to explore these potential applications, assuming landmarks can be accurately detected. Our actual acoustic landmark detection accuracy, despite increasing over time, has not reached a practical level yet.

3.2 Measures of Information in Acoustic Frames

An acoustic landmark is an instantaneous event that serves as a reference time point for the measurement of spectrotemporal cues widely separated in time and frequency. For example, in the paper that first defined landmarks, Stevens proposed classifying distinctive features of the landmark based on the onsets and offsets of formants and other spectrotemporal cues up to 50ms before or 150ms after the landmark [23]. The 200ms spectrotemporal dynamic context proposed by Stevens is comparable to the 165ms spectrotemporal dynamic context computed for every frame by the ASR system [64]. Most ASR systems use acoustic features that are derived from frames 25ms long with a 10ms skip because human speech is quasi-stationary for this short period [65]. Because spectral dynamics commu-

nicate distinctive features, however, ASR systems since 1981 ([66]) have used dynamic features; since deep neural nets (DNNs) began gaining popularity, the complexity of the dynamic feature set in each frame has increased quite a lot, with consequent improvements in ASR accuracy. This trend not only applies to stacking below 100ms. With careful normalization, features like TRAPs [67] with temporal windows equal or longer than 500ms continue to demonstrate accuracy improvement.

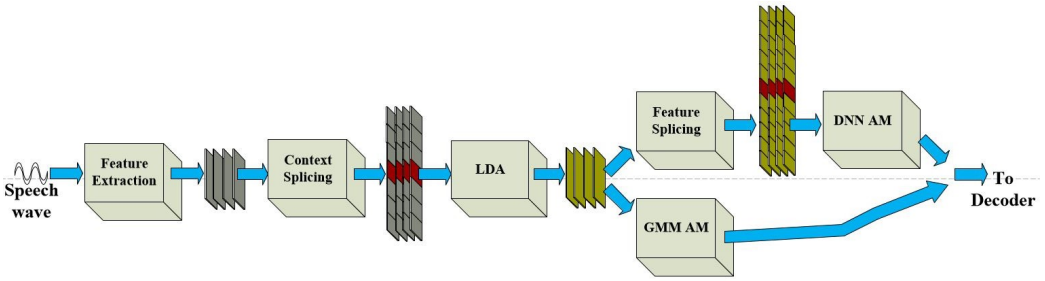


Figure 3.1: Stacking of feature frames before the scoring process for DNN AM (top path) and GMM AM (bottom path). The dark gray, red and green rectangles indicate frames and stacks of frames.

Experiments reported in this paper are built on a baseline described by [64], and schematized in Fig. 3.1. In this system, MFCCs are computed once every 10ms, with 25ms windows (dark gray rectangles in Fig 3.1). In order to include more temporal context, we stack adjacent frames, three preceding and three succeeding, for a total of seven frames (a total temporal span of 85ms). These are shown in Fig 3.1 as the longer, segmented dark gray rectangles, with the red segments representing the center frames of each stack. The seven-frame stack is projected down to 40 dimensions using linear discriminant analysis (LDA). For input to the DNN but not the GMM, LDA is followed by speaker adaptation using mean subtraction and feature-space maximum likelihood linear regression (fMLLR), additional context is provided by a second stacking operation afterwards, in which LDA-transformed features, represented by yellow rectangles, are included in stacks of 9 frames (for a total temporal span of 165ms), as represented by the top path in Figure 3.1. It is believed that the reason features spanning longer duration improve ASR accuracy is that long lasting features capture coarticulation better, including both neighboring-phone transitions and longer-term coarticulation. The dynamics of the tongue naturally cause the articulation of one phoneme to be reflected in the transitions into and out of neighboring phonemes, over a time span of perhaps

70ms. Longer-term coarticulation, spanning one or more syllables, can occur when an intervening phoneme does not require any particular placement of one or more articulators. For example, öhman [68] demonstrated that the tongue body may transit smoothly from a vowel to the next without apparently being constrained by the presence of several intervening consonants.

3.2.1 Re-Weighting Frames

HMM-based ASR searches the space of all possible state sequences for the most likely state sequence given the observations. Conventional decoding procedure assumes equal weights for the acoustic likelihoods of all speech frames. However, paying much more attention to informative frames, such as frames anchored at landmarks, may be beneficial to the performance of beam searching according to the suggestion from acoustic landmark theory. We modified the log likelihood of a state sequence $\mathbf{s}$ given the observations $\mathbf{o}$ in the following equation,

$$\log P\left(\mathbf{s}|\mathbf{o}\right) \propto \sum_{t=1}^{T} w_t \cdot \log P\left(o_t|s_t\right) + \log P\left(s_t|s_{t-1}\right) \tag{3.1}$$

where s_t and o_t represents the senone[1] state and observed acoustic feature vector at time t, respectively. The transition probability between senone states is denoted as $P\left(s_t|s_{t-1}\right)$, and the logarithm of the emitting probability $\log P\left(o_t|s_t\right)$ is scaled by the weighting factor w_t. If the speech frame vector o_t occurs at a landmark, w_t is assigned to a larger value[2], otherwise, w_t is set to 1.

Tuning an optimal value of w_t is in a similar way of finding a suitable scaling value of acoustic likelihoods over language models. If the frame over-weighted is a frame that can differentiate the correct state better, the error rate will drop. In contrast, if the likelihood of a frame is divided evenly across states, or even worse, is higher for the incorrect state, then over-weighting this frame will mislead the decoder and increase chances of error. For this reason, over-weighting landmark frames is a good measure to tell how meaningful landmark frames are compared to the rest of the frames. If the landmarks are indeed more significant, we should observe a reduction in the PER for the system over-weighting the landmark.

[1] Senones are represented as either monophone states or clustered triphone states.

[2] We defined it as over-weighting in the following context if w_t is greater than 1.

3.2.2 Dropping Frames

The wide temporal windows used in modern ASR, as mentioned in the beginning of Section 3.2, are highly useful to landmark-based speech recognition: all of the dynamic spectral cues proposed by [23] are within the temporal window spanned by the feature vector of a frame centered at the landmark; therefore it may be possible to correctly identify the distinctive features of the landmark by dropping all other frames, and keeping only the frame centered at the landmark. Our different frame dropping heuristics modify the log probability of a state sequence by replacing the likelihood $P\left(o_t|s_t\right)$ with an approximation function $f\left(\cdot\right)$. In terms of log probabilities, Equation (3.1) becomes

$$\log P\left(\mathbf{s}|\mathbf{o}\right) = \sum_{t=1}^{T} \log f\left(P\left(o_t|s_t\right), t\right) + \log P\left(s_t|s_{t-1}\right) \tag{3.2}$$

The class of optimizations considered in this paper involve a set of functions $f(P(o_t|s_t), t)$ parameterized as:

$$f\left(P\left(o_t|s_t\right), t\right) = \begin{cases} R\left(\mathbf{o}, t\right), & \text{if } g\left(t\right) = 1 \\ P\left(o_t|s_t\right), & \text{otherwise} \end{cases} \tag{3.3}$$

The *method of replacement* is characterized by R, and the frame-dropping function by $g(t)$. This work considers multiple methods to verify that the finding with respect to landmarks is independent of the replacement method. The four possible settings of the $R(\mathbf{o}, t)$ function are as follows:

$$R(\mathbf{o}, t) \in \begin{cases} R_{\text{Copy}}(\mathbf{o}, t) = P(o_{t'}|s_{t'}), \quad t' = \max_{\tau \leq t,\ g(\tau)=0} \tau \\ R_{\text{Fill\_0}}(\mathbf{o}, t) = 1 \\ R_{\text{Fill\_const}}(\mathbf{o}, t) = \left(\prod_{t=1}^{T} P(o_t|s_t)\right)^{1/T} \\ R_{\text{Upsample}}(\mathbf{o}, t) = \exp\left(\sum_{t':g(t')=0} h(t-t') \log P(o_t|s_t)\right) \end{cases} \tag{3.4}$$

In other words, the *Copy* strategy copies the most recent observed value of $P(o_t|s_t)$, the *Fill_0* strategy replaces the log probability by 0, the *Fill_const* strategy replaces the log probability by its mean value, and the *Upsample* strategy replaces it by an interpolated value computed by interpolating (using interpolation filter $h(t)$) the log probabilities that have been selected for retention. The *Upsample* strategy will only be used if the frame-dropping function is periodic, i.e., if frames are

downsampled by a uniform downsampling rate.

The *pattern of dropped frames* can be captured by the indicator function g, which is true for frames that we want to drop. Experiments will test two landmark-based patterns: *Landmark-drop* drops all landmark frames ($g(t) = 1$ if the frame contains a landmark), and *Landmark-keep* keeps all landmark frames ($g(t) = 1$ only if the frame does *not* contain a landmark). In the case where landmark information is not available, the frame-dropping pattern may be *Regular*, in which $g(t) = \delta(\ t \bmod K\ )$ indicating that every K-th frame is to be dropped, or it may be *Random*, in which case the indicator function is effectively a binary random variable set at a desired frame dropping rate. As we will demonstrate later, to achieve a specific function and dropping ratio, we can sometimes combine output of different g functions together by taking a logical inclusive OR to their output.

If acoustic landmark frames contain more valuable information than other frames, it can be expected that experiment setups that retain the landmark frames should out-perform other patterns, while those that drop the landmark frames should under-perform, regardless of the *method of replacement* chosen.

3.3 Hypotheses

This paper tests two hypotheses. The first is that a window of speech frames (in this case 9 frames) centered at a phonetic landmark has more information than windows centered elsewhere – this implies that over-weighting the landmark-centered windows can result in a reduction in PER. The second hypothesis states that keeping landmark-centered windows rather than other windows causes little PER increment, and that dropping a landmark-centered window causes greater PER increment as opposed to dropping other frames. In the study we focused on PER as opposed to Word Error Rate (WER) for two reasons. First, the baseline Kaldi recipe for TIMIT reports PER. Second, this study is oriented towards speech acoustics; focusing on phones allow us to categorize and discuss the experiment and results in better context.

In order to test these hypotheses, a phone boundary list from the TIMIT speech corpus [56] was obtained, and the landmarks were labeled based on the phone boundary information. Table 3.1 briefly illustrates the types of landmarks and their positions, as defined by the TIMIT phone segments. This marking procedure is shared by [17,31,44]. It is worth mentioning that this definition disagrees with that

Table 3.1: Landmark types and their positions for acoustic segments. **Fc** and **Fr** are closure and release for fricatives; **Sc** and **Sr** are closure and release for Stops; **Nc** and **Nr** are closure and release for nasals; **V** and **G** are vowel pivot and glide pivot; 'start', 'middle', and 'end' denote three positions across acoustic segments.

Manner	Landmark	Observation in Spectrogram
Vowel	V: middle	maximum in low- and mid-frequency
Glide	G: middle	minimum in low- and mid-frequency
Fricative	Fc: start, Fr: end	amplitude discontinuity occurs when consonantal constriction is formed or released
Affricate	Sr,Fc: start, Fr: end	
Nasal	Nc: start, Nr: end	
Stop	Sc: start, Sr: end	

of [41]. Lulich claims that there is no landmark in the center of Vowel and Glide; instead, a formant-subglottal resonance crossing, which is known to sit between the boundaries of [-Back] and [+Back] vowels, contains a landmark. Frames marked as landmark are of interest. To test hypothesis 1, landmark frames are over-weighted. To test hypothesis 2, either non-landmark or landmark frames are dropped.

3.4 Experimental Methods

Our experiments are performed on the TIMIT corpus. Baseline systems use standard examples distributed with the Kaldi open source ASR toolkit[3]. Specifically, the GMM-based baseline follows the configurations in the distributed `tri2` configuration in the Kaldi TIMIT example files[4]. The clustered triphone models are trained using maximum likelihood estimation of features that have been transformed using linear discriminant analysis and maximum likelihood linear transformation. For the DNN baseline, speaker adaptation is performed on the features, and nine consecutive frames centered at the current frame are stacked as inputs to the DNN, as specified in the distributed `tri4_nnet` example. Respectively, the two systems achieved PER of 23.8% (GMM) and 22.6% (DNN) without any modification.

We performed a 10-fold cross validation (CV) over the full corpus, by first combining the training and test sets, and creating 10 disparate partitions for each test condition. The gender balance was preserved to be identical to the canonical test set for each test subset, while the phonetic balance was approximately the same

[3] `http://kaldi-asr.org/`

[4] `https://github.com/kaldi-asr/kaldi/tree/master/egs/timit/s5`

but not necessarily identical. This is in order to improve the significance of our PER numbers. The TIMIT corpus is fairly small and the phone occurrence of some phones, or even phone categories, in the test set is lower than ideal. Conducting cross validation on the full set allows us partially address this issue.

For the control experiments of our tests, all configurations of feature extraction and decoding process are retained the same as the baseline. In this case, fair comparisons are guaranteed, and we can fully reveal the effects of our methods in the Acoustic Model (AM) scoring process.

3.5 Experimental Results

Experimental results examining the two hypotheses proposed above will be presented in this section. We will present the results of over-weighting the landmark frames first. Evaluation of frame dropping will be presented second, and includes several phases. In the first phase, a comparison of different *methods of replacement* is presented, to provide the reader with more insight into these methods before they are applied to acoustic landmarks. In the second phase, we will then leverage our findings to build a strategy that both drops non-landmark frames, and over-weights landmark frames, using the best available *pattern of dropped frames* and *method of replacement*. We open source the code used to carry out the following experiments online [5].

3.5.1 Hypothesis 1: Over-weighting Landmark Frames

Figure 3.2 illustrates the PER of the strategy of over-weighting the landmark frames during the decoding procedure, and how it varies with the factor used to weight the AM likelihood of frames centered at a landmark. The PER for GMM-based models drops as the weighting factor increases until the factor is 1.5; increasing the weighting factor above 1.5 causes the PER to increase slightly. When the factor is increased to greater than 2.5, the PER increases at a higher slope. Similar trends can be found for DNN models, yet in this case the change in PER is non-concave and spans a smaller range. If landmark frames are under-weighted, or over-weighted by a factor of 1.5 or up to 2.0, PER increases. Over-weighting landmark frames by

[5] `https://github.com/dihe2/kaldi/tree/master/egs/timit/s5`

a factor of 3.0 to 4.0 reduces PER. In this experiment, Wilcoxon tests [69] have been conducted, through Speech Recognition Scoring Toolkit (SCTK) 2.4.10[6], and tests concluded the difference to be insignificant.

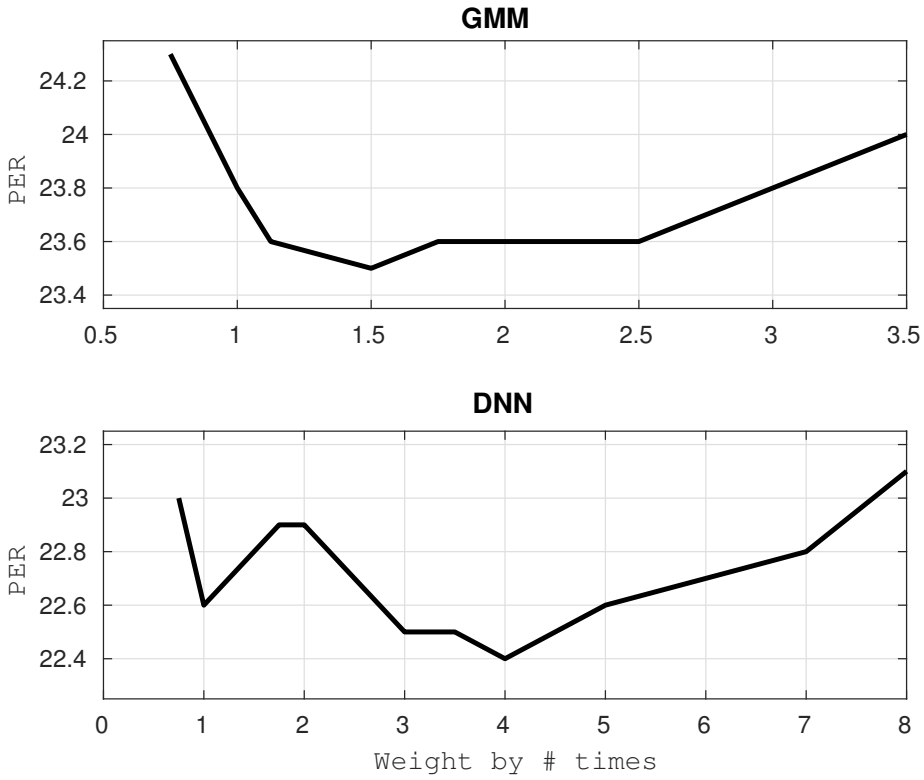


Figure 3.2: Over-weighting landmark frames for GMM and DNN.

3.5.2 Methods of Replacement for Dropped Frames

Figure 3.3 compares the performance of three *methods of replacement*: *Copy*, *Fill_0* and *Fill_const* when a *Regular* frame dropping pattern is used. Results show that *Fill_0* and *Fill_const* suffer very similar PER increments as the percentage of frames dropped is increased, while *Copy* shows a relatively smaller PER increment for drop rates of 40% or 50%. As for the comparison between acoustic models, DNN-based models outperform GMM-based at all drop rates. Notably, the *Copy* approach synergizes well with DNN models, and is able to maintain low PER increments even up to 75% drop rate; this finding is similar to findings reported in papers from [62].

[6] `https://www.nist.gov/itl/iad/mig/tools`

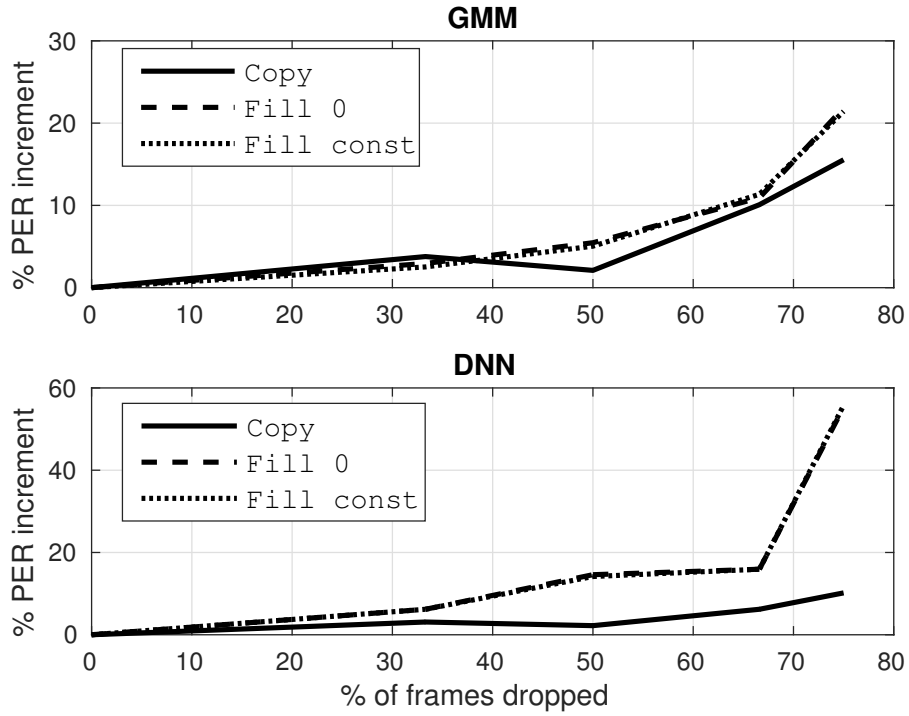


Figure 3.3: Comparison of Different Methods of Frame Replacement (*Copy*, *Fill_0* and *Fill_const*) assuming a *Regular* pattern of frame replacement.

Figure 3.4 compares the performance between two *patterns of dropping frames* – *Regular*, *Random*. In both of these the *Copy* method for replacement was used. We also provide for comparison, the *Regular* pattern, but using an *Upsample* replacement method. This scheme uses a 17-tap anti-aliasing FIR filter. The method that offered the lowest phone error rate increment is obtained using a *Regular* pattern with a *Copy* replacement scheme. Results show that *Regular-Copy* outperforms other methods by a large margin in terms of PER increment independent of which AM is used.

3.5.3 Hypothesis 2: Dropping Frames with Regards to Landmarks

At the beginning of this section, experiments that test hypothesis 2 directly are described. The focus is to subject the ASR decoding process to frames missing acoustic likelihood scores, and see how the decoding error rate changes accordingly. Obviously we are interested in using the presence vs. absence of an acoustic landmark as a heuristic to choose the frames to keep or drop. To quantify the importance of the information kept vs. the information discarded, dropping strategies

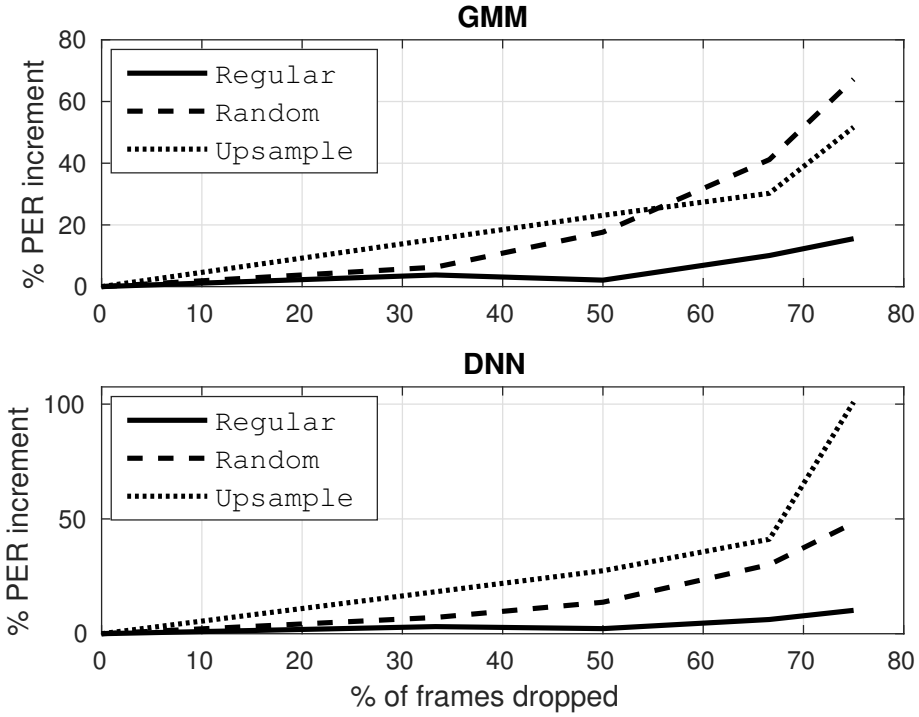


Figure 3.4: Comparison of Different Patterns of Dropping Frames assuming *Copy* (*Regular* and *Random*) and *Interpolation through low-pass filtering* (*Upsample*) method of replacement.

(*Landmark-keep* and *Landmark-drop*) are compared to the non-landmark-based *Random* strategy. Notice the *Regular* strategy has been shown to be more effective than *Random* (e.g., in Fig. 3.4), however, to make the PER result meaningful, the same number of frames should be dropped across different patterns being compared. When we keep only landmarks (*Landmark-keep*) or drop only landmarks (*Landmark-drop*), the percentage of frames dropped can not be precisely controlled by the system designer: it is possible to adjust the number of frames retained at each landmark (thus changing the drop rate), but it is not possible to change the number of landmarks in a given speech sample. Therefore, precisely adjusting the drop rate to meet a different pattern is not practical. Depending on the test set selected, the portion of frames containing landmarks ranges from 18.5% to 20.5%. As opposed to *Random*, *Regular* does not give us the ability to select a drop rate that exactly matches the drop rate of the *Landmark-drop* or *Landmark-keep* strategies. Therefore, it is not covered in the first 2 experiments. However, in the 3rd experiment, we will compare a frame dropping strategy using landmark as heuristic against *Regular* dropping. But that experiment will serve a slightly different purpose.

As in the over-weighting experiment, two types of frame replacement are tested. The *Fill_0* strategy is an exact implementation of hypothesis 2: when frames are dropped, they are replaced by the least informative possible replacement (a log probability of zero). Figure 3.3 showed, however, that the *Copy* strategy is more effective in practice than the *Fill_0* strategy; therefore these two strategies are tested using a landmark-based frame drop pattern. Figure 3.3 showed that the *Fill_const* strategy returns almost identical results to *Fill_0*, so it is not separately tested here.

Experiment results are presented for both the TIMIT default test split, and for cross-validation (CV) using the whole corpus. The baseline implementation is as distributed with the Kaldi toolkit. Since no frames are dropped, it returns the lowest PER. However, likelihood scoring for the baseline AM will require more computation when compared to a system that drops frames. For CV we report the mean relative PER increment ($\Delta\text{PER} = 100 \times (\text{modified PER} - \text{baseline PER})/(\text{baseline PER})$), with its standard deviation in parentheses, across all folds of CV. Every matching pair of frame-drop systems (*Landmark-keep* versus *Random*) is tested using a two-sample t-test [70], across folds of the CV, in order to determine whether the two PER increments differ. During the t-test, we assume PER numbers from different folds are samples of a random variable. The two-sample t-test intends to find out whether the random variables representing

PER for different setups (*Landmark-keep* versus *Random*) have the same mean.

Keeping or Dropping the Landmark Frames

Table 3.2 illustrates the changes in PER increment that result from a *Landmark-keep* strategy (score only landmark frames) versus a *Random* frame-drop strategy set to retain the same percentage of frames. For each test set, we count the landmark frames separately and match the drop rate exactly between the *Landmark-keep* and *Random* strategy. In all cases, the *Landmark-keep* strategy has a lower PER increment. A Wilcoxon test, other then the two-sample t-test, has been conducted on the default test set; the differences between all pairs but the DNN *Fill0* pair is significant on this test.

Table 3.2: PER increments for scoring landmark frames only compared to randomly dropping similar portion of frames. CV stands for cross validation; if the two increments differ, then the lower of the two is marked with either $*$ ($p < 0.05$) or $**$ ($p < 0.001$).

Acoustic model	GMM				DNN			
Test regime	Default		CV Mean (Stdev)		Default		CV Mean (Stdev)	
Metric	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)
Baseline	23.8	0.0	22.8	0.0	22.7	0.0	20.8	0.0
Fill_0								
Landmark-keep	36.1	51.7	33.4	46.5(1.34)**	49.6	118.5	49.7	139(10.3)*
Random	42.3	77.7	42.1	84.6 (8.35)	50.9	124.2	52.8	154 (14.8)
Copy								
Landmark-keep	35.2	47.7	32.3	41.5(1.08)**	29.4	29.3	26.9	29.3(0.653)**
Random	44.0	84.9	44.1	93.5 (0.734)	38.4	69.3	37.6	80.9 (0.942)

For the next experiment we inverted the setup: instead of keeping only landmark frames, we drop only landmark frames (call this the *Landmark-drop* strategy). Table 3.3 compares the PER increment of a *Landmark-drop* strategy to the increment suffered by a *Random* frame drop strategy with the same percentage of lost frames. The *Landmark-drop* strategy always return higher PER. However, only for the GMM setup *Copy* did we obtain a significant p value during cross validation. The p values for other setups range from 0.13 to 0.17. Again, a Wilcoxon test, other than the two-sample t-test, has been conducted on the default test set, with the conclusion that only the GMM *Copy* pair demonstrated significant difference.

The results in Table 3.2 demonstrate that keeping landmark frames is better than keeping a random selection of frames at the same drop rate, in all but one of the tested comparison pairs. The results in Table 3.3 demonstrate that random selection tends to be better than selectively dropping the landmark frames, though

Table 3.3: PER increments for dropping Landmark frames during scoring compared to randomly dropping a similar portion of frames (CV stands for cross validation)

Acoustic model	GMM				DNN			
Test regime	Default		CV Mean (Stdev)		Default		CV Mean (Stdev)	
Metric	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)	PER (%)	PER Inc (%)
Baseline	23.8	0.0	22.8	0.0	22.7	0.0	20.8	0.0
Fill_0								
Landmark-drop	25.6	7.56	24.0	5.33(1.36)	24.2	6.61	23.1	11.1(1.58)
Random	24.1	1.26	23.4	2.68 (1.23)	23.6	3.96	22.4	7.53 (1.24)
Copy								
Landmark-drop	25.6	7.5	24.1	5.83(0.873)*	24.3	7.1	22.1	6.44(0.836)
Random	24.6	3.3	23.1	1.14 (0.948)	23.6	4.0	21.6	3.85 (0.760)

the difference is only significant in one of the four comparison pairs. These two findings support the hypothesis that frames containing landmarks are more important than others. However, the PER increment in some setups are very large, indicating the ASR might no longer be functioning under stable conditions.

Using Landmark as a Heuristic to Achieve Computation Reduction

Methods in Tables 3.2 and 3.3 compared the *Landmark-keep*, *Landmark-drop*, and *Random* frame drop strategies. Table 3.4 illustrates PER increment (%) for the *Landmark-keep* and *Regular* frame-dropping strategies. In this experiment, we are no longer directly testing Hypothesis 2. Instead, we are trying to achieve high frame dropping rate subject to low PER increment. As dropped frames need not be calculated during the acoustic model scoring procedure, a high dropping ratio can benefit the ASR by reducing computational load. The strategy leveraging landmark information is a hybrid strategy: on top of a standard *Regular* strategy, it keeps all landmark frames and over-weights the likelihoods of these frames as in 3.5.1. For each acoustic model type (GMM vs. DNN), three different percentage rates of frame dropping are exemplified. In each case, we select a *Regular* strategy with high dropping rate, modify it to keep the landmark frames, measure the percentage of frames dropped by the resulting strategy, then compare the result to a purely *Regular* frame-drop strategy with a similar drop rate. The baseline *Regular* strategies have three standard drop rates: 33.3% (one out of three frames dropped, uniformly), 50% (one out of two frames dropped), and 66.7% (two out of three frames dropped). Table 3.4 highlights results for one of the setups in bold, as that setup achieves a very good trade off between high dropping ratio and low PER increment.

Table 3.4: PER increments comparison between Landmark-keep and Regular drop strategies for GMM and DNN.

	Copy	Default		Cross Validation			
		Drop Rate%	PER Inc%	Drop Rate%	PER Inc%	Inc STD%	Inc pVal
GMM	Land	41.0	1.26	44.4	1.84	0.0133	0.962
	Reg	33.3	3.78	33.3	1.81	0.0119	
	Land	54.2	2.94	54.1	2.86	0.0140	0.598
	Reg	50	2.1	50	2.58	0.00780	
	Land	64.3	12.1	65.0	8.10	0.0182	0.159
	Reg	66.7	10.1	66.7	6.91	0.0181	
DNN	Land	41.0	0.44	44.4	1.84	0.0115	0.0011
	Reg	33.3	3.98	33.3	4.20	0.0153	
	Land	**54.2**	**0.44**	**58.4**	**1.90**	**0.167**	**0.0029**
	Reg	50	2.21	50	4.12	0.0115	
	Land	64.2	3.08	69.0	5.86	0.0121	0.0391
	Reg	66.7	6.17	66.7	7.04	0.0160	

Table 3.5: PER increments (%) for Landmark-keeping strategy for DNN with dropping rate near 54.2% and over-weighting factor near 4 times

Drop Rate%	Over-weighting Factor		
	3.5	4	4.5
52.1	1.42	0.84	0.93
54.2	0.88	0.44	0.88
56.3	0.62	0.40	0.40

As we can see, for DNN acoustic models, the *Landmark-keep* strategy results in lower error rate increment than a *Regular* strategy dropping a similar number of frames. Wilcoxon tests demonstrated a statistically significant difference at all three drop rates. For GMM acoustic models, avoiding landmarks does not seem to return a lower error rate. In fact, the error rate is higher for 2 out of 3 different drop rates. The highlighted case in Table 3.4 is intriguing because it the PER increment is so low, and this row will therefore serve as the basis for further experimentation in the next section. In this setup for DNN, over 50% of the frames were dropped, but the PER only increased by 0.44%. This result seems to support the hypothesis that landmark frames contain more information for ASR than other frames, but in Table 3.4, this row has the appearance of an anomaly, since the error increment is so small. In order to confirm that this specific data point is not a special case, we conducted additional experiments with very similar setups. The results for these additional experiments are presented in Table 3.5.

Additional results presented in Table 3.5 are obtained through applying an

over-weighting factor close to 4, which is the optimal value found for DNNs in Figure 3.2. The first and third rows in this table randomly keep or drop a small number of non-landmark frames, in order to obtain drop rates of 52.1% and 56.3% respectively. Since the selection is random, multiple runs of the experiment result in different PER for the same drop rate; therefore we repeated each experiment 10 times and reported the mathematical mean. Since there is a level of randomness in these results, we do not intend to evaluate our hypotheses on these data; rather, the goal of Table 3.5 is merely to confirm that the highlighted case in Table 3.4 is a relatively stable result of its parameter settings, and not an anomaly. Since good continuity can be observed across nearby settings, results in Table 3.5 lend support to the highlighted test case in Table 3.4.

3.6 Discussion

Results in Section 3.5.1 tend to support hypothesis 1. However, the tendency is not statistically significant. The tendency is consistent for the GMM-based system, for all over-weighting factors between 1.0 and 3.0. Similar tendencies appeared for over-weight factors between 3.0 and 5.0 for DNN-based system.

Experiments in Section 3.5.2 tested different non-landmark-based frame drop strategies, and different methods of frame replacement. It was shown that, among the several strategies tested, the *Regular-Copy* strategy obtains the smallest PER. There is an interesting synergy between the frame-drop strategy and the frame-replacement strategy, in that the PER of a 50% *Regular-Copy* system (one out of every two frames dropped) is even better than that of a 33% *Regular-Copy* system (one out of every three frames dropped). This result, although surprising, confirms a similar finding reported by [71]. We suspect that the reason may be relevant to the regularity of the 50% drop rate. When we drop 1 frame out of every 2 frames, the effective time span of each remaining frame is 20ms, with the frame extracted at the center of the time span. Dropping 1 frame out of every 3 frames, on the other hand, results in an effective time span per frame of 15ms, but the alignment of each frame's signal window to its assigned time span alternates from frame to frame.

It is worth mentioning that our definition of acoustic landmarks differs from that of [41] – specifically, Lulich claims that there is no landmark in the center of Vowel and Glide. Instead, formant-subglottal resonance crossing, which is known to sit between the boundaries of [-Back] and [+Back] vowels, contains a landmark.

It is possible that an alternative definition of landmarks might lead to better results.

We can also observe that GMM and DNN acoustic models tend to perform differently in the same setup. For example, for GMM, randomly dropping frames results in a higher PER than up-sampling; this is not the case for DNN models. Results also demonstrate that DNN models perform quite well when frames are missing. A PER increment of only 6% occurs after throwing away $2/3$ of the frames. GMM models tend to do much worse, especially when the drop rate goes up.

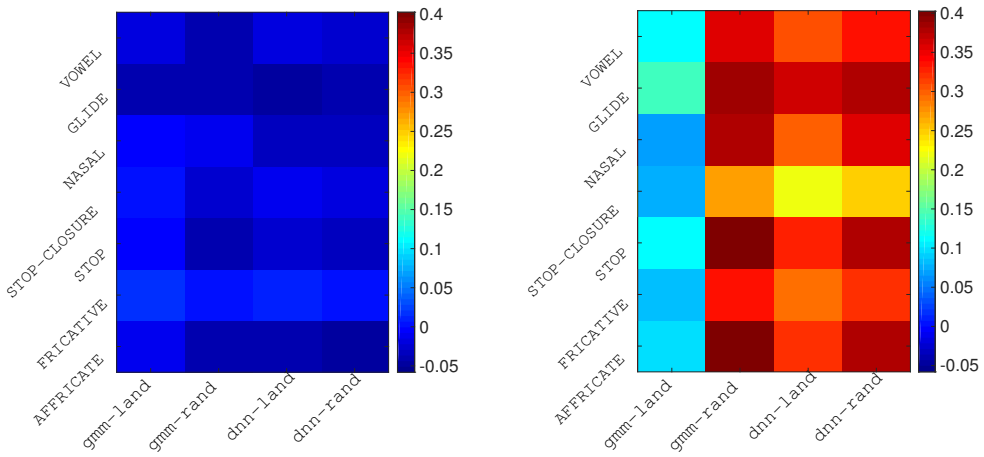
All experiments on DNN tend to support the strategy to avoid dropping landmarks. However, the 2 test cases covered in Table 3.3 lack statistical confidence. Scoring only the landmark frames (the *Landmark-keep* strategy) out-performs both *Random* and *Regular* frame-drop-strategies. On the other hand, if landmark frames are dropped (the *Landmark-drop* strategy), we obtain higher PER when compared to randomly scoring a similar number of frames.

We find, at least for ASR with DNN acoustic models, that landmark frames contain information that is more useful to ASR than other frames. In the most striking case, the highlighted result in Table 3.4 indicates that it is possible to drop more than 54% of the frames but only observe a 0.44% increment in the PER compared to baseline (PER increases from 22.7 to 22.8). We conclude, for DNN-based ASR, that experiments support hypothesis 2 (with statistically significant differences in two out of the three comparisons). In comparison, we failed to find support for hypothesis 2 in GMM-based ASR.

3.6.1 How Landmarks Affect the Decoding Results

Having proven that the *Landmark-keep* strategy is more effective than a *Random* or *Regular* drop strategy, we proceeded to investigate the resulting changes in the rates of insertion, deletion and confusion among phones. We compared the normalized increment of each type of error, separately, when the confusion matrices of the baseline system are subtracted from the confusion matrices of the *Landmark-keep* and *Random* frame-drop systems. Fig. 3.5 compares the normalized error increment, of different types of errors, for the *Landmark-keep* and *Random* strategies. The numbers reported in the figure are normalized error increment. They are calculated using error increment divided by the occurrence of each kind of phone. We use this measure to reflect the increment ratio while avoiding having to deal with situations

that could lead to division by zero.



(a) *Insertion errors* **(b)** *Deletion errors*

Figure 3.5: The normalized error increment for a) insertion errors and b) deletion errors (y-axis represent different manners of articulators and x-axis represent different systems)

Overall, dropping frames causes a minor reduction to the phone insertion rate, while the phone deletion rate significantly worsens. We suspect that after dropping frames, the decoder is less effective at capturing transitions between phones, resulting in correctly detected phones spanning over other phones. In Figure 3.5b we can see that the *Landmark-keep* strategy is more effective than the *Random* strategy, since it returns a lower deletion rate increment. We believe this is because the landmark contains sufficient acoustic information about each phone to force it to be recognized. However, we do not know why the GMM-`Landmark-keep` strategy is less effective at preventing phone deletions than the DNN-`Landmark-keep` strategy. A possible reason might be that more frames were stacked together in the splicing process for the DNN than for the GMM [64]. If we do consider providing landmarks as extra information to ASR, in order to reduce computation load for example, the difference between GMM and DNN models should be considered.

3.7 Conclusions

Phones can be categorized using binary distinctive features, which can be extracted through acoustic cues anchored at acoustic landmarks in the speech utterance. In this work, we proved through experiments for DNN-based ASR systems operating on MFCC features, on the TIMIT corpus, using both the default and cross validation train-test splits, that frames containing landmarks are more informative than others. We proved that paying extra attention to these frames can potentially compensate for accuracy lost when dropping frames during Acoustic Model likelihood scoring. We leveraged the help of landmarks as a heuristic to guide frame dropping during speech recognition. In one setup, we dropped more than 54% of the frames while adding only 0.44% to the Phone Error Rate. This demonstrates the potential of landmarks for computational reduction for ASR systems with DNN acoustic models. We conclude that a DNN-based system is able to find a nearly-sufficient summary of the entire spectrogram in frames containing acoustic landmarks, in the sense that, if computational considerations require one to drop 50% or more of all speech frames, one is better off keeping the landmark frames than keeping any other tested set of frames. GMM-based experiments return mixed results, but results for the DNN are consistent and statistically significant: landmark frames contain more information about the phone string than frames without landmarks.

CHAPTER 4

MULTI-TASK LEARNING WITH ACOUSTIC LANDMARKS FOR LOW-RESOURCED LANGUAGE

Furui first demonstrated that the identity of both consonant and vowel can be perceived from the C-V transition; later, Stevens proposed that acoustic landmarks are the primary cues for speech perception, and that steady-state regions are secondary or supplemental. Acoustic landmarks are perceptually salient, even in a language one doesn't speak, and it has been demonstrated that non-speakers of the language can identify features such as the primary articulator of the landmark. These factors suggest a strategy for developing language-independent automatic speech recognition: landmarks can potentially be learned once from a suitably labeled corpus and rapidly applied to many other languages. This chapter proposes enhancing the cross-lingual portability of a neural network by using landmarks as the secondary task in multi-task learning (MTL). The network is trained in a well-resourced source language with both phone and landmark labels (English), then adapted to an under-resourced target language with only word labels (Iban). Landmark-tasked MTL reduces source-language phone error rate by 2.9% relative, and reduces target-language word error rate by 1.9%-5.9% depending on the amount of target-language training data. These results suggest that landmark-tasked MTL causes the DNN to learn hidden-node features that are useful for cross-lingual adaptation.

4.1 Introduction

In the early 1980s, Furui [26] demonstrated that the identity of both consonant and vowel can be perceived from a 100ms segment of audio extracted from the C-V transition; in 1985, Stevens [23] proposed that acoustic landmarks are the primary cues for speech perception, and that steady-state regions are secondary or supplemental. Acoustic landmarks produce enhanced response patterns on the mammalian auditory nerve [72], and it has been demonstrated that non-speakers of

a language can identify features such as the primary articulator of the landmark [73]. Automatic speech recognition (ASR) systems have been proposed that depend completely on landmarks, with no regard for the steady-state regions of the speech signal [74], and such systems have been demonstrated to be competitive with phone-based ASR under certain circumstances. Other studies have proposed training two separate sets of classifiers, one trained to recognize landmarks, another trained to recognize steady-state phone segments, and fusing the two for improved accuracy [44] or for reduced computational complexity [12]. It has been difficult to build cross-lingual ASR from such systems, however, because very few of the world's languages possess large corpora with the correct timing of consonant release and consonant closure landmarks manually coded. In this chapter we propose a different strategy: we propose to use reference landmark labels in only one language (the source language). A landmark detector trained in the source language is ported to the target language in two ways: (1) by automatically detecting landmark locations in target language test data, and (2) by using landmark detection as a secondary task for the purpose of training a triphone state recognizer that can be more effectively ported cross-lingually. The neural network is trained with triphone state recognition as its primary task; landmarks are introduced as a secondary task, using the framework of multi-task learning (MTL) [75].

MTL has shown the ability to improve the performance of speech models, especially those based on neural networks [18, 76–78]. MTL is a mechanism for reducing generalization error. A single-task neural net is provably optimal, for large enough training datasets: as the size of the training dataset goes to infinity, if the number of hidden nodes is set equal to the square root of the number of training samples, the difference between the network error rate and the Bayes error rate goes to zero [79]. MTL is useful when the training dataset is too small to permit zero-error learning [76], or when the training dataset and the test dataset are drawn from slightly different probability distributions (e.g., different languages). In either case, MTL proposes training the network to perform two tasks simultaneously. The secondary task is not important during test time, but if the network is forced to perform the secondary task during training, it will sometimes learn network weights (and consequently, hidden layer activation functions) that are either (1) less prone to over-fitting on the training data than a single-task network, or (2) generalize better from the distribution of the training data to the distribution of the test data. Landmark detection could potentially be an ideal secondary task for automatic speech recognition (ASR; Fig 4.1), since it detects instantaneous events that are

informative to phone recognition. Because landmarks have been demonstrated to correlate with non-linguistic perceptual signals (e.g., enhanced response on the auditory nerve [72]) and because features of a landmark can be classified by non-speakers of the language [73], it is possible that the secondary task of landmark detection and classification will force a neural net to learn weights that are more useful for cross-language ASR adaptation [17] than those of a single-task network. These characteristics are especially helpful for under-resourced languages: in an under-resourced language, training data may be limited, e.g., there may be little or even no transcribed speech. A Landmark-based system trained on a well-resourced language might be adapted to an under-resourced language, thus improving ASR accuracy in the under-resourced language.

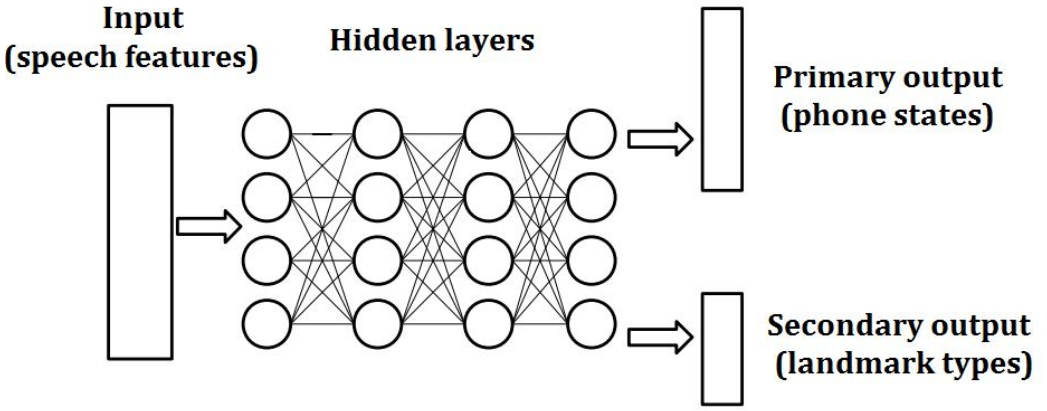


Figure 4.1: MTL Neural Network Jointly Trained on Phone States and Landmark Types

The key contributions of this work are experimental findings supporting the hypothesis that landmark-task MTL reduces the word error rate of a cross-lingually ported ASR. After we review some background in Sec. 4.2, key methodology and techniques used to apply the Landmark theory to MTL are explained in Sec. 4.3. Results are presented in Sec. 4.4, and the chapter concludes in Sec. 4.5.

4.2 Background

Before we talk about our methodology, we would like to briefly review MTL as a neural network training method and talk about the under-resource corpus we used in this study.

4.2.1 Multi-task Learning

Multi-task Learning (MTL) [75] has shown the ability to improve statistical model performance by jointly training a single model for multiple purposes. The multiple tasks in MTL share the same input, but generate multiple outputs predicting likelihoods for a primary and one or more secondary tasks. When the multiple tasks are related but not identical, or (in the ideal case) complementary to each other, MTL models offer better generalization from training to test corpus [76]. A number of works [76–78] have proved MTL to be effective on speech processing tasks. Among them [78] proved MTL effective at improving model performance for under-resourced ASR.

When we conduct MTL, for the same input x, we prepare two sets of labels. The label l_i^{ph} specifies the phone or triphone state associated with a frame, while l_j^{la} encodes the presence and type of acoustic landmark. The network is trained in order to minimize, on the training data, a multi-task error metric as shown in Eq. 4.1, where $P_i^{ph}(x)$ $(1 \leq i \leq C^{ph})$ is the probability of monophone or triphone state i at frame x as estimated by the neural network, $P_j^{la}(x)$ $(1 \leq j \leq C^{la})$ is the probability of landmark label j at frame x as estimated by the network, and α is a trade-off value we use to weight the two sets of labels. We sweep through a small list of candidate α's to find the value that returns the best result on development test data.

$$\mathcal{L}_{mtl} = (1-\alpha)\sum_{i=1}^{C^{ph}}(l_i^{ph}\log(P_i^{ph}(x))) + \alpha\sum_{j=1}^{C^{la}}(l_j^{la}\log(P_j^{la}(x))) \tag{4.1}$$

4.2.2 The Iban Corpus

The under-resourced language studied in this chapter is Iban [80]. Iban is a language spoken in Borneo, Sarawak (Malaysia), Kalimantan and Brunei. The Malay phone set is similar to English, e.g., the two languages have the same inventory of stop consonants and affricates; Malay also has a relatively transparent orthography, in the sense that the pronunciation of a word is usually well predicted by its written form. If Iban orthography is as transparent as Malay, and if its phone set is as similar to English, then it is possible that a landmark detector trained on English may perform well in Malay. Iban is also selected for these experiments because of the recent release of an Iban training and test corpus with particularly good

quality control [80]. The Iban corpus contains 8 hours of clean speech from 23 speakers. Seventeen speakers contributed $6.8h$ of training data, and the test-set contains $1.18h$ of data from 6 speakers. The language model was trained on a $2M$-word Iban news dataset using SRILM [81].

4.3 Methods

We trained an ASR on the TIMIT corpus using the methods of multi-task learning (Sec. 4.2.1), using the detection and classification of landmarks (Sec. 4.3.1) as a secondary task. The same ASR is then adapted cross-lingually to the Iban corpus (Sec. 4.2.2)

4.3.1 Defining and Marking Landmarks

Landmark definitions in this chapter, listed in Table 4.1, are based primarily on those of [43], with small modifications. Modifications include the elimination of the +33% and -20% offsets after the beginning or before the end of some phones, reported in [43] and [40], in favor of the simpler definitions in Table 4.1.

Table 4.1: Landmark types and their positions for acoustic segments, where 'c', and 'r' denote consonant closure, and release; 'start', 'middle', and 'end' denote three positions across acoustic segments, respectively.

Manner of Articulation	Landmark Type and Position
Vowel	V: middle
Glide	G: middle
Fricative	Fc: start, Fr: end
Affricate	Sr,Fc: start, Fr: end
Nasal	Nc: start, Nr: end
Stop Closure	Sc: start, Sr: end

We extracted landmark training labels by referencing the TIMIT human annotated phone boundaries. An example of the labeling is presented in Fig 4.2. This example from [12] illustrates the labeling of the word "symposium"[1]. The figure is generated using Praat [82].

[1] selected from audio file: `TIMIT/TRAIN/DR1/FSMA0/SX361.WAV`

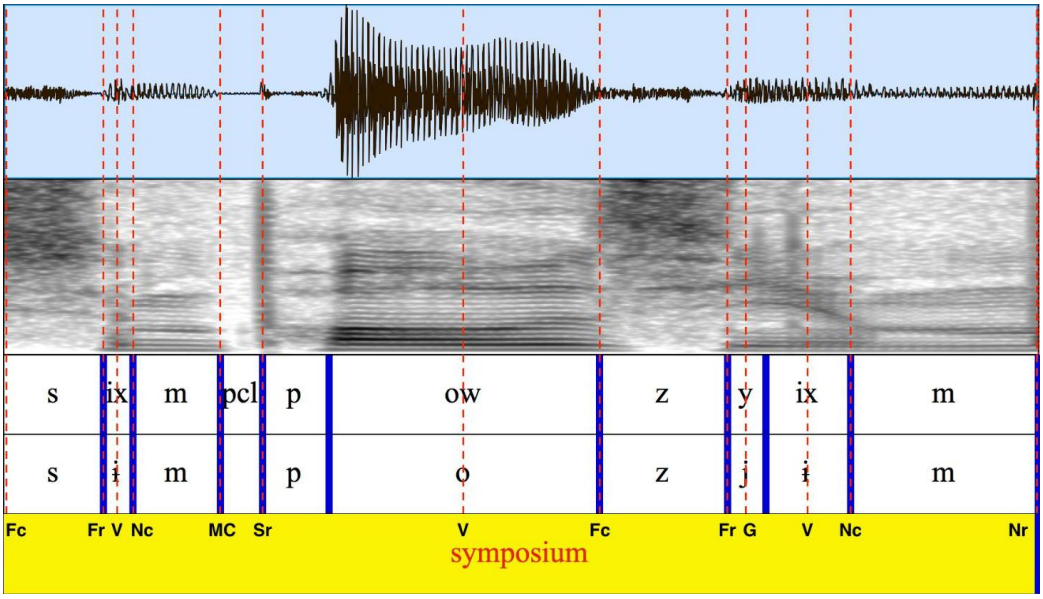


Figure 4.2: Acoustic landmark labels for the pronunciation of word "symposium".

Landmarks are relatively infrequent compared to phone-state-labeled speech frames: every frame has a phone label, but fewer than 20% of frames have a Landmark label. Because of the sparsity of Landmark-labeled frames, we explored different ways to adjust the Landmark labels to achieve the best MTL performance. We found, expanding the range of a Landmark to include the nearby 2 frames returns the highest accuracy for the primary task.

To further address the imbalance among different Landmark classes, the training criterion was computed using a weighted sum of training data, with weights inversely proportion to the class support.

4.3.2 Cascading the MTL to Iban

After we trained a landmark detector on TIMIT, we ran the detector on Iban. The English-trained landmark detector output is used to define reference labels for the secondary task of the Iban acoustic model MTL. An example of the detector output on an arbitrary utterance[2] in Iban is given in Fig 4.3. We found that the results are good at outlining fricative landmarks. The detector can also find stop closure landmarks near the correct locations, but with less precision than the fricative landmarks. The performance on vowel and glide landmarks is only fair: the

[2] `iban/data/wav/ibm/003/ibm_003_049.wav`

detector often mixes up the two classes, and incorrectly labels sonorant consonants as vowels.

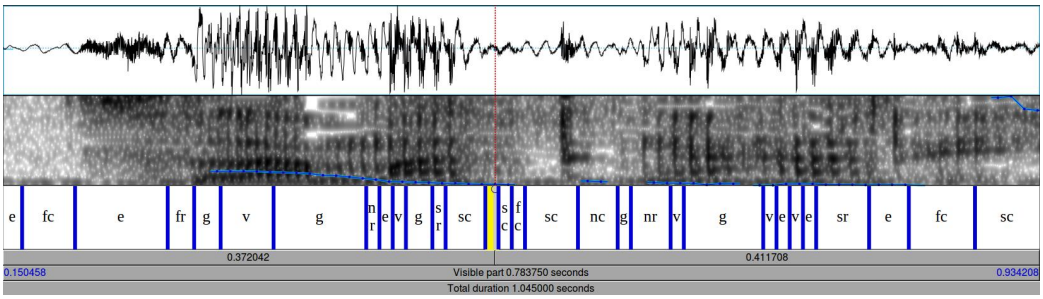


Figure 4.3: Landmark Detection Result on Iban for utterance ibm_003_049, pronouncing **selamat tengah ari** (**s-aa-l-a-m-a-t t-aa-ng-a-h a-r-i** in Iban phone set). Transcription labels: e=empty (no Landmark); fr, fc, sr, sc, nr, nc, v, g are as in Table 1.

When applying the landmark detector to Iban, we are concerned with the error generated by the detector. The automatically detected landmark labels are treated as ground truth for MTL in landmark-task MTL in Iban, therefore it is possible that erroneously detected landmarks may mis-lead the network training. To minimize the effect of these mistakes, we introduce an extra weighting factor in the MTL training criterion based on the confidence of the landmark detector output, as shown in Eq. 4.2.

$$\mathcal{L}_x = (1 - \alpha c_x) \sum_{i=1}^{C^{ph}} (l_i^{ph} \log(P_i^{ph}(x))) + \alpha c_x \sum_{j=1}^{C^{la}} (l_j^{la} \log(P_j^{la}(x))) \tag{4.2}$$

where c_x is a confidence value derived based on the landmark detector output for feature frame x based on Eq 4.3.

$$c_x = P_m^{la\_de}(x) - \frac{1}{C^{la} - 1} \sum_{k=1, k \neq m}^{C^{la}} (P_k^{la\_de}(x)) \tag{4.3}$$

where $P_i^{la\_de}(x)$ is the softmax output for landmark class i. The class index $m = \arg\max_i^{C^{la}} P_i^{la\_de}(x)$, which is also the index for the class the landmark detector predicted.

The intuition behind this extra layer of weighting is to assign a penalty, during training of the ASR, that is proportional to our certainty of its error. If the detector is not confident separating the output class from other classes, then we reduce the

loss it generates in the MTL process.

We experimented with multiple ways to initialize the landmark detector and the phone recognizer in the second language. We found that using a network trained through MTL in TIMIT to initialize the MTL network in the second language yields the best results. We found the technique marginally but consistently outperforms other initializations including DBN.

4.4 Results

We extracted an acoustic feature vector using the same algorithm and parameters as [77]. The acoustic model (AM) is a deep neural network with 4 hidden, fully-connected layers, 2048 nodes/layer. The same features and network structure were used for both the landmark detector, the MTL model and the baseline. The baseline is initialized using a DBN [83]. No speaker adaptation is used in any of the ASR systems in this chapter.

Results are reported in Table 4.2 for both English (TIMIT) and Iban. TIMIT results are reported to indicate the performance of Landmark-based MTL in the source language, prior to cross-language adaptation.

On development test sets in both corpora, the value $\alpha = 0.2$ returned the lowest error rate (with little variability in the range $0.1 \leq \alpha \leq 0.3$), and was therefore used for evaluation. The landmark detector achieves 80.11% frame-wise accuracy in validation. Phone error rate (PER) was reasonably good: 20.6% for the baseline system, and 20.0% for the MTL system, as compared to 22.7% for the open-source Kaldi tri4_nnet recipe.

Decoding results for Iban are reported using Word Error Rate (WER), because the Iban corpus is distributed with automatic but not manual phonetic transcriptions. The comparison between PER in TIMIT and WER in Iban permits us to demonstrate that Landmark-based MTL can benefit PER in a source language (English), and WER in an adaptation target language (Iban). Triphone-based ASR trained without MTL on TIMIT, then adapted to Iban, achieves 18.4% WER; a system that is identical but for the addition of landmark-task MTL can achieve 17.93% WER. Neither system includes speaker adaptation, and therefore neither system is better than the 17.45% state of the art WER for this corpus[3] with the

[3] https://github.com/kaldi-asr/kaldi/blob/master/egs/iban/s5/RESULTS

same language model.

Table 4.2: Decoding Error Rate for mono-phone (Mono) and tri-phone (Tri) on TIMIT and Iban.

Corpus	AM	Baseline	MTL	MTL w/ Confid
TIMIT (PER)	Mono	24.6	24.2	-
	Tri	20.6	20.0	-
Iban-full (WER)	Mono	24.62	24.22	24.18
	Tri	18.40	18.03	17.93
Iban-25% (WER)	Mono	28.87	27.97	27.64
	Tri	21.31	20.70	20.63
Iban-10% (WER)	Mono	31.16	28.49	28.48
	Tri	25.12	23.64	23.57

As we can see in Table 4.2, in all cases, regardless of AM and corpus, the ASR system jointly trained with landmark and phone information returns lower error rate. The setups "Iban-25%" and "Iban-10%" train the AM on only 25% (100 minutes) and 10% (40 minutes) of the training data uniformly selected at random from the Iban training set (maintaining speaker and gender ratio), but evaluates the error rate on the full test set. As the amount of training data decreases, the benefits of MTL increase. When only 10% of training data is available, simulating a very low resource case, MTL reduces the word error rate by the greatest margin: 8.7% for monophone ASR and 6.17% for triphone ASR. Weighting the MTL loss according to confidence results in a small but consistent error rate reduction. All systems use the same language model, and all systems use acoustic models with the same network architecture and feature set; the error rate change we observe is caused entirely by the use of landmark-task MTL.

4.5 Discussion

This demonstrates that landmark-task MTL results in a neural network that can be more effectively ported cross-lingually. As the amount of training data in the under-resourced language is reduced (from 400 minutes to 100 or 40 minutes), the benefits of landmark-task MTL increase. In addition, introducing a loss weighting according the landmark detector confidence seems to reduce the effect of landmark detector error as it consistently produces lower error rate.

While a cross-language Landmark detector provides useful information complementary to the orthographic transcription, visual inspection indicates that a cross-language landmark detector is not as accurate as a same-language landmark detector. Future work, therefore, will train a more accurate landmark detector, using recurrent neural network methods that do not depend on human-annotated phone boundaries, and that can therefore be more readily applied to multi-lingual training corpora.

CHAPTER 5

PRONUNCIATION ERROR IDENTIFICATION ON CHINESE LEARNING

This chapter explores a novel approach of identifying pronunciation errors for the second language (L2) learners based on the landmark theory of human speech perception. Earlier works on the selection method of distinctive features and the likelihood-based "goodness of pronunciation" (GOP) measurement have gained progress in several L2 languages, e.g. Dutch and English. However, the improvement of performance is limited due to error-prone automatic speech recognition (ASR) systems and less distinguishable features. Landmark theory posits the existence of quantal nonlinearities in the articulatory-acoustic relationship, and provides a basis of selecting landmark positions that are suitable for identifying pronunciation errors. By leveraging this English acoustic landmark theory, we propose to select Mandarin Chinese salient phonetic landmarks for the Top-16 frequently mispronounced phonemes by Japanese (L1) learners, and extract features at those landmarks including mel-frequency cepstral coefficients (MFCC) and formants. Both cross validation and evaluation are performed for individual phonemes using support vector machine with linear kernel. Experiments illustrate that our landmark-based approaches achieve higher micro-average f1 score significantly than GOP-based methods.

5.1 Introduction

Pronunciation error identification, as an essential technology in computer assisted pronunciation training systems that provide an effective way of enhancing the speaking skills for the second language (L2) learners, attracts considerable attention from research communities of speech signal processing and applied linguistics.

With the advance in automatic speech recognition (ASR) research, solutions that identify pronunciation errors have made great progress recently. These systems typically detect segmental (phone level) mispronunciations from L2 learner's read

speech using an ASR decoder, and pinpoint salient pronunciation errors, such as insertions, substitutions, or deletions of specific pronunciation units [84]. More specifically, two types of ASR-based mispronunciation detection techniques have been widely applied. *Rule-based* approach uses extended pronunciation confusion networks that include both canonical pronunciations and their mispronounced variants [85–87]; *Confidence-based* approach measures the similarity between native speaker's canonical pronunciation and its corresponding realization by L2 learners [88–90]. ASR utilizes hidden Markov models (HMMs) to capture temporal information of phones, however, HMMs are still not powerful enough to discriminate sounds that are spectrally similar and differ mainly in duration [91]. For example, HMMs are not quite suitable to distinguish fricatives from plosive release segments since the difference of these two sounds is subtle in the amplitude envelope [92].

Therefore, another line of this research, as illustrated in this chapter, will cast pronunciation error identification as a binary classification task that improves discrimination power by detecting distinctive feature errors known to occur with high frequency. Acoustic landmark theory [60] by exploiting quantal nonlinear articulatory-acoustic relationships provides a basis of selecting distinctive features that are suitable for speech recognition [44]. We will leverage this theory further for the task of identifying pronunciation errors.

5.2 Related Works

The factors that cause high-frequency errors differ for L2 learners from different native language backgrounds. For example, the single biggest pronunciation problem for Spanish-speaking learners of English is that Spanish does not have a distinction between short and long vowels [93], while Japanese-speaking learners can mitigate the affects of vowel duration [94].

Acoustic cues that distinguish error minimal pairs include standard ASR features, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP), but also include many specialized cues that have been found to be useful for particular binary contrasts. Voice onset time (VOT) features are proposed to detect Mandarin Chinese (L2) phonetic errors of aspirated consonants (/p/, /t/, /k/) pronounced by Japanese native speakers [95]. Rate of Rise (ROR) values that are calculated by determining the derivative of the logarithm of root-mean-square

energy succeed to discriminate the voiceless velar plosive /k/ from the voiceless velar fricative /x/ in Dutch, since the release of the burst of the plosive causes an abrupt rise in amplitude [91]. As for the mispronounced Dutch vowels, formants (F1-F3) have been used with ASR-based confidence measures are exploited further [96]. Goodness of speech prosody (GOSP) has also been defined [97] in terms of several features including F0, duration, parameters of Fujisaki model, rPVI, and nPVI.

Among these feature representations of mispronounced sounds, statistical models are also explored with the purpose of selecting distinctive features. Stouten et al [98] applied artificial neural network models to extract distinctive features from MFCC features in the context of learning English. Lee et al [99] tried to identify error sounds by leveraging features that are learned from neural networks in an unsupervised manner. Hacker et al [100] achieved promising performance based on top-15 distinctive features using the AdaBoost algorithm for the task of detecting English errors made by German children.

Recent application of landmark-based distinctive features in ASR motivated researchers to further explore their utility in pronunciation error detection problems. Quantal nonlinearities in articulatory-acoustic relations provide a theoretical basis for selecting distinctive features, complementary to the empirical foundations of most L2 research [60]. Acoustic landmark theory, first described in [23], has been successfully applied in identifying English pronunciation errors produced by Korean speakers [101, 102]. This guidance of selecting distinctive features is probably suitable for a larger pool of languages other than English only. In the context of Chinese learning as a foreign language, Zhang et al [103, 104] developed a Mandarin Chinese distinctive features system based on the knowledge of acoustics and physiology. Wang et al demonstrated the capability of discrimination between several phonemes on that system by comparing the parameters of perceptual linear prediction (PLP), MFCC and linear prediction cepstral coefficients (LPCC) [105]. However, determining the acoustic landmark positions that best represent categorical phonological distinctions remains a difficult problem, since the acquisition of this knowledge requires large scale experiments of human speech perception [105]. The lack of this knowledge hinders the progress of the application on identifying pronunciation errors.

In this chapter, we provide two alternative methods for selecting acoustic landmark positions in L2 Chinese. First, we directly mapped well-founded English landmark theory into Mandarin Chinese, since there exists similar phonetic char-

acteristics between these two languages; second, we define Chinese landmark positions and corresponding distinctive features and acoustic cues by analyzing a large scale corpus of pronunciation error pairs.

5.3 Description of Data

This large scale corpus of Chinese (L2) speech, referred to as BLCU inter-Chinese speech corpus [106], has collected data from more than 100 speakers. Each speaker read a sentence from 301 daily use sentences. This corpus consists of 64,190 phonemes and 4,631 utterances. The continuous speech of 17 Japanese native speakers (8 males and 9 females) has been phonetically annotated at segment level. The annotators are 6 post-graduate students majoring in phonetics, divided into two groups. The speech data were annotated twice independently by the two groups, with each annotator labeling a continuous 200 utterances on a rotating basis. The speech data were manually transcribed and were automatically aligned into phonetic segments of "initials" and "finals" with human transcriptions using HTK Speech Recognition Toolkit [107]. The absolute agreement (in percentage of matching values) between annotators ranges from 77.0% to 84.3% with the average agreement 80.7%. The correlation coefficients are computed for the phoneme based mispronunciation rates for the two groups with the average correlation ratio 0.78. 65 kinds of pronunciation error tendencies (PETs) based on articulation-placement and articulation-manner are annotated to represent general erroneous articulation tendencies, including raising, lowering, advancing, backing, lengthening, shortening, centralizing, rounding, spreading, labio-dentalizing, laminalizing, devoicing, voicing, insertion, deletion, stopping, fricativizing, nasalizing, and retroflexing [106].

In this study, we only considered consistent utterances where two annotators are in agreement, and select Top-16 frequent PETs to perform experiments on identifying pronunciation errors, while ignoring other very rare PETs. These 16 PETs were divided into four categories: spreading, backing, shortening and laminalizing. The final corpus consists of 7,837 phones (error: 1,524, correct: 6,313). The error rate across 16 phones ranges from 6% to 44.1%.

5.4 Methodology

In the context of pronunciation error identification, *confidence-based* approaches still maintain better performance than *rule-based* methods due to the "coverage of confusion networks" trade-offs [108]. The goodness of pronunciation (GOP) [89] algorithm is probably the most widely used measure in this scope. Our baseline for testing landmark-based pronunciation error detections is GOP using deep neural network triphone acoustic models trained on our large scale corpus [109].

5.4.1 Goodness of Pronunciation (GOP)

The aim of GOP is to provide a confidence score for each phone in a speech utterance. Given the orthographic transcription and acoustic models that determine the likelihood $P(O^q|q)$ where O^q denotes the acoustic segment aligned with phone $q \in Q$, the GOP score can be calculated by normalizing the log likelihood ratio of phone p compared to its strongest competitor over the number of frames $NF(p)$ in phone p,

$$GOP(p) = \left| \log \left(\frac{P(O^p|p)}{\max_{q \in Q} P(O^q|q)} \right) \right| / NF(p) \tag{5.1}$$

We applied maximum mutual information (MMI) estimation to adapt acoustic models using Japanese native speaker's speech. The numerator and denominator in Eq. (5.1) are computed by forced alignment with orthographic transcription and an unconstrained phone loop, respectively.

5.4.2 Acoustic Landmark Theory

Stevens proposed [25, 60] four different candidate landmark locations for English, including vowel peak landmark, oral closure landmark, glide valley landmark in glide-like consonants, and oral release landmark. These four landmark categories were proposed by Stevens to be language-universal, but our studies of Mandarin Chinese suggest other signal events that may have a better claim to be both perceptually salient and phonologically distinctive.

In the task of pronunciation error identification, we could explore error pairs from the development corpus in order to define the acoustic landmark positions and distinctive features. We conduct speech perception experiments in collaboration with experts at BLCU Department of Linguistics, and discovered the distinctive

Chinese landmark positions for 16 phones with high error frequencies in the corpus. As an alternative to these perceptually-based Chinese landmark candidates, we find correspondences of articulatory-manner and articulatory-place between English and Mandarin Chinese after applying Stevens theory. Table 5.1 lists the landmark positions signaling 16 Chinese phonemes according to these two different methods of landmark definition.

Table 5.1: Acoustic landmark positions obtained by Chinese phonetics and English phonetics theory. Phone symbols are IPA (*pinyin* in parens). The fraction number denotes the relative time stamp in the duration.

Phone	Chinese Landmark	English Landmark
ʂ (sh)	following vowel	fricative: start, end
ʤ (zh)	coda of consonant	affricate: start, end
ʧ (ch)	onset of consonant	affricate: start, end
ɕ (x)	following vowel	fricative: start, end
ʤ (j)	following vowel	affricate: start, end
an (an)	onset (14/30) of vowel	nasal: start,end
y (v)	onset of vowel	vowel: middle
aŋ (ang)	onset (14/30) of vowel	nasal: start, end
iŋ (ing)	onset (17/30) of vowel	nasal: start, end
u (u)	onset of vowel	vowel middle
f (f)	onset of consonant	fricative: start, end
əŋ (eng)	onset of vowel	vowel: start, end
ʧ (q)	onset, nucleus, coda	affricate: start, end
k (k)	following vowel	stop: start
ɻ (r)	whole consonant	fricative: start, end
uɔ (uo)	onset, nucleus, coda	glide: middle

5.5 Experiments and Results

We compared acoustic landmark features (see Table 5.1) with GOP-based features in this section using 10-fold stratified cross validation. Evaluations were then performed using a held-out test set.

5.5.1 Setup

In all experiments, MFCC appended by its acceleration, delta coefficients, and C0 coefficients are extracted. Cepstral mean normalization is applied to compensate long-term spectral effects[1]. Formants[2] (F1 - F5) are computed from the signal up to 5500Hz since all test-takers are female. A Hamming window of 25ms was used to chunk short-term stationary signals as frames, and the default frame rate is 10ms. However, many Chinese phones have short durations, and many segments therefore contain less than four frames (e.g. /u/ and /i/ often have one frame). To address the issues of insufficient number of frames, MFCC were recomputed using a frame rate adjusted as necessary between 2ms and 10ms. 20% of the whole corpus was held out as a test set that holds the same proportions of class labels.

Taking into account the imbalanced nature of the training set, we applied support vector machine (SVM) with linear kernel, and assigned weights inversely proportional to the class frequencies as suggested in [21].

5.5.2 Cross Validation

Figure 5.1 illustrates the micro-average f1 scores for each individual phones over six different features. Red bars denote GOP baseline models that hold reasonable performance for most of the phones except /aŋ/(ang) and /k/(k). In comparison to GOP baseline, all acoustic cues at landmarks outperform GOP measure significantly except for the phones /f/(f), /tɕ/(ch), and /ɻ/(r) due to large overlaps of error bars.

From the comparisons between MFCC features at Chinese landmarks (blue) and English landmarks (green), we observed that for the nasal phones /aŋ/(ang), /iŋ/(ing), and /an/(an) with backing errors, English landmarks outperform Chinese. According to the landmarks definition for these phones as shown in Table 5.1, Chinese landmarks fall on the onset of vowels while English landmark considered beginning and end of the consonant, which seems to be a better position for discriminating tongue backing errors in both the vowel and the consonant. For the fricative phones /ʃ/(sh), /ɕ/(x), and /dʑ/(j), Chinese landmarks located at the following vowel perform worse than English consonant-boundary landmarks, despite the perceptually salient vowel difference that co-occurs with the consonant distinction in Chinese.

[1] HCopy config parameter: `MFCC_0_A_D_Z`

[2] `http://www.fon.hum.uva.nl/praat/`

Formants features (light and cyan) that are literally expected to disambiguate vowels, seem not to contribute for discriminations for all phones except for the phones /iŋ/(ing) and /ɻ/(r). The aspirated stop phone /k/(k) expresses an interesting pattern that the Chinese landmark (the following vowel) achieves a better f1 score than English (only considering the start of the stop release segment). Aggregating all features together as shown in yellow bars made limited improvements particularly in the case that both Chinese and English landmarks compensate with each other, e.g. fricative phone /ʂ/(sh).

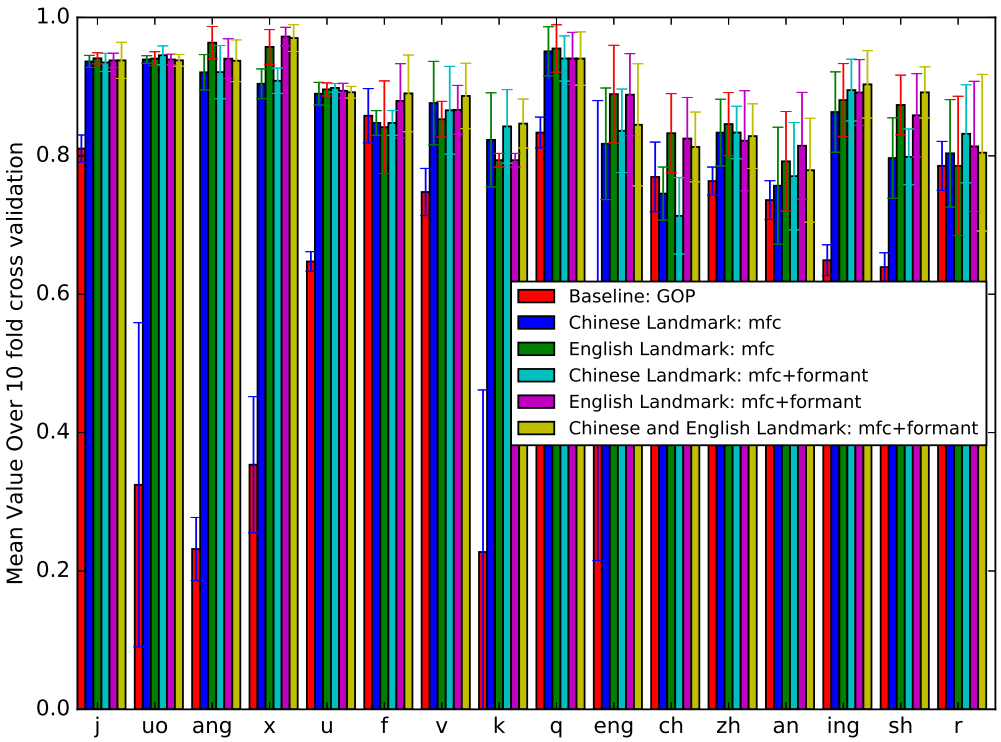


Figure 5.1: 10-fold cross validation performances of GOP baseline and acoustic features at Chinese and English landmarks. Y-axis shows the micro-average values of f1 scores for each individual phones. The sequence of phones is sorted by the error percentage in the training set.

The observation in Figure 5.1 indicates that acoustic cues beyond the basic MFCC may still be redundant or irrelevant for the classification tasks. For example, formants at the onset of consonant could be irrelevant feature for the affricative phone /tʂ/(ch), and may need to be removed. While for the affricative /dʑ/(j) and glide type /uɔ/(uo), the performance remains unchanged after applying MFCC and formants features at Chinese and English landmarks. Besides of the landmark

positions and corresponding acoustic cues, various frame rates are also applied on 16 phones. We empirically choose the best acoustic cues for each individual phones based on the best micro-average f1 scores (see Table 5.2). Seven phones /aŋ/(ang), /y/(v), /tɕ/(q), /əŋ/(eng), /ʈʐ/(zh), /iŋ/(ing), ʂ(sh) can achieve the best performance by using large frame rate (10ms), while the smaller frame rate (4ms) are suitable two phones /uɔ/(uo) and /k/(k).

Table 5.2: Best acoustic cues selected for individual phones.

Phone	FrameRate (ms)	Landmark	AcousticCues	F1score
dʑ (j)	6	Chn+Eng	mfc+formant	0.949
uɔ (uo)	4	Eng	mfc+formant	0.945
aŋ (ang)	10	Eng	mfc+formant	0.967
ɕ (x)	6	Eng	mfc+formant	0.977
u (u)	8	Chn	mfc+formant	0.890
f (f)	8	Chn+Eng	mfc+formant	0.902
y (v)	10	Chn+Eng	mfc+formant	0.887
k (k)	4	Chn	mfc	0.872
tɕ (q)	10	Eng	mfc	0.970
əŋ (eng)	10	Eng	mfc	0.908
ʈʂ (ch)	8	Eng	mfc	0.861
ʈʐ (zh)	10	Eng	mfc+formant	0.855
an (an)	6	Eng	mfc	0.844
iŋ (ing)	10	Eng	mfc+formant	0.919
ʂ (sh)	10	Eng+Chn	mfc+formant	0.902
r (r)	8	Chn	mfc+formant	0.832

5.5.3 Acoustic Cues for Evaluation

Cross validation experiments demonstrate that the performances for individual phones under micro-average f1 score were highly correlated with the combinations of landmark positions, frame rates, and acoustic cues. We continued to explore the generalization power of these models on our 20% held-out test data. In this study, the best frame rate for each phone was frozen as shown in Table 5.2, and six models including GOP baseline were evaluated. In the context of identifying pronunciation errors, L2 learners expect to receive more feedbacks of pinpointing salient errors rather than false alarms. Receiver Operating Characteristic (ROC) metric that formulates the relationship between true positive rate (TPR) and false

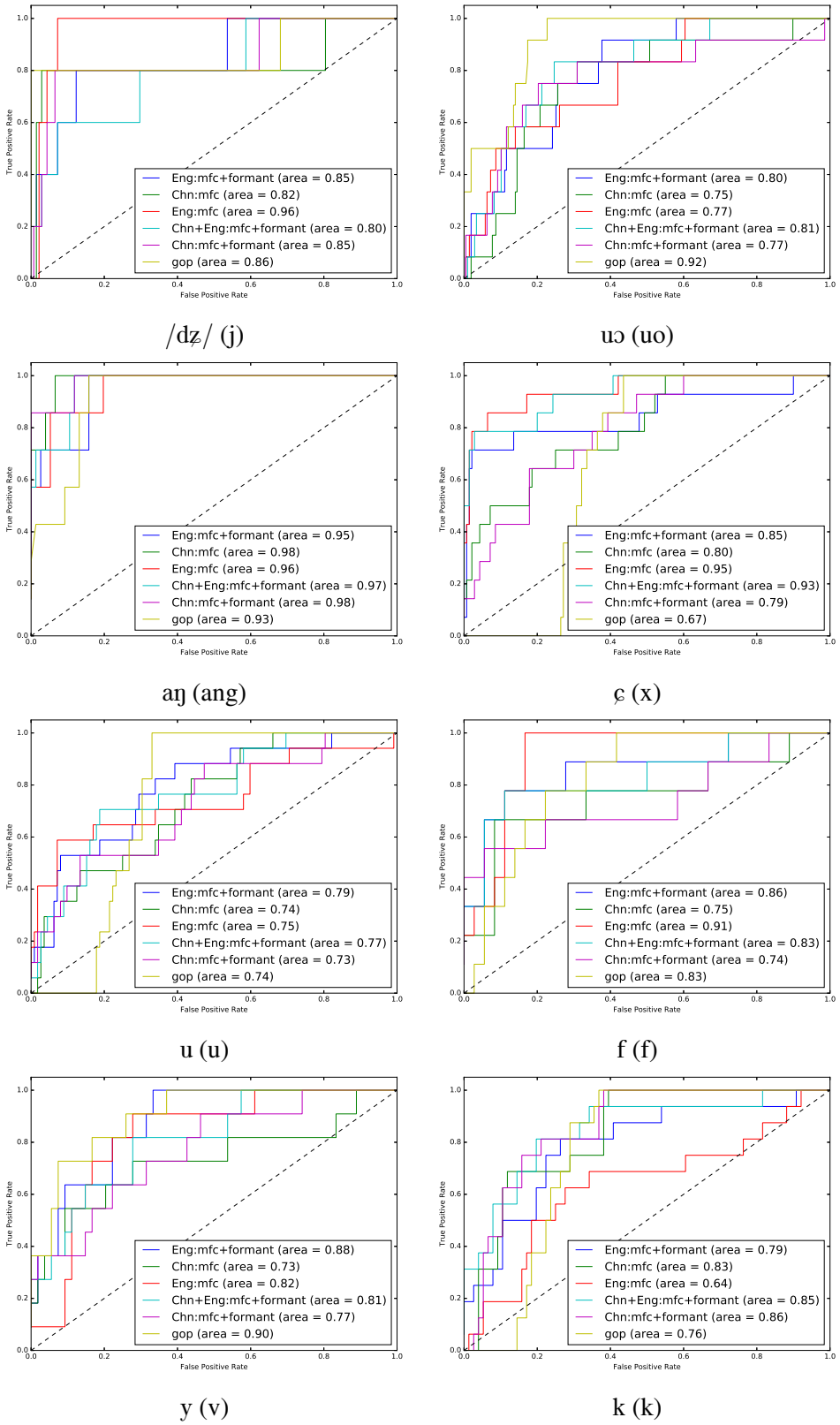


Figure 5.2: ROC curves of evaluations on held-out test set for each phone.

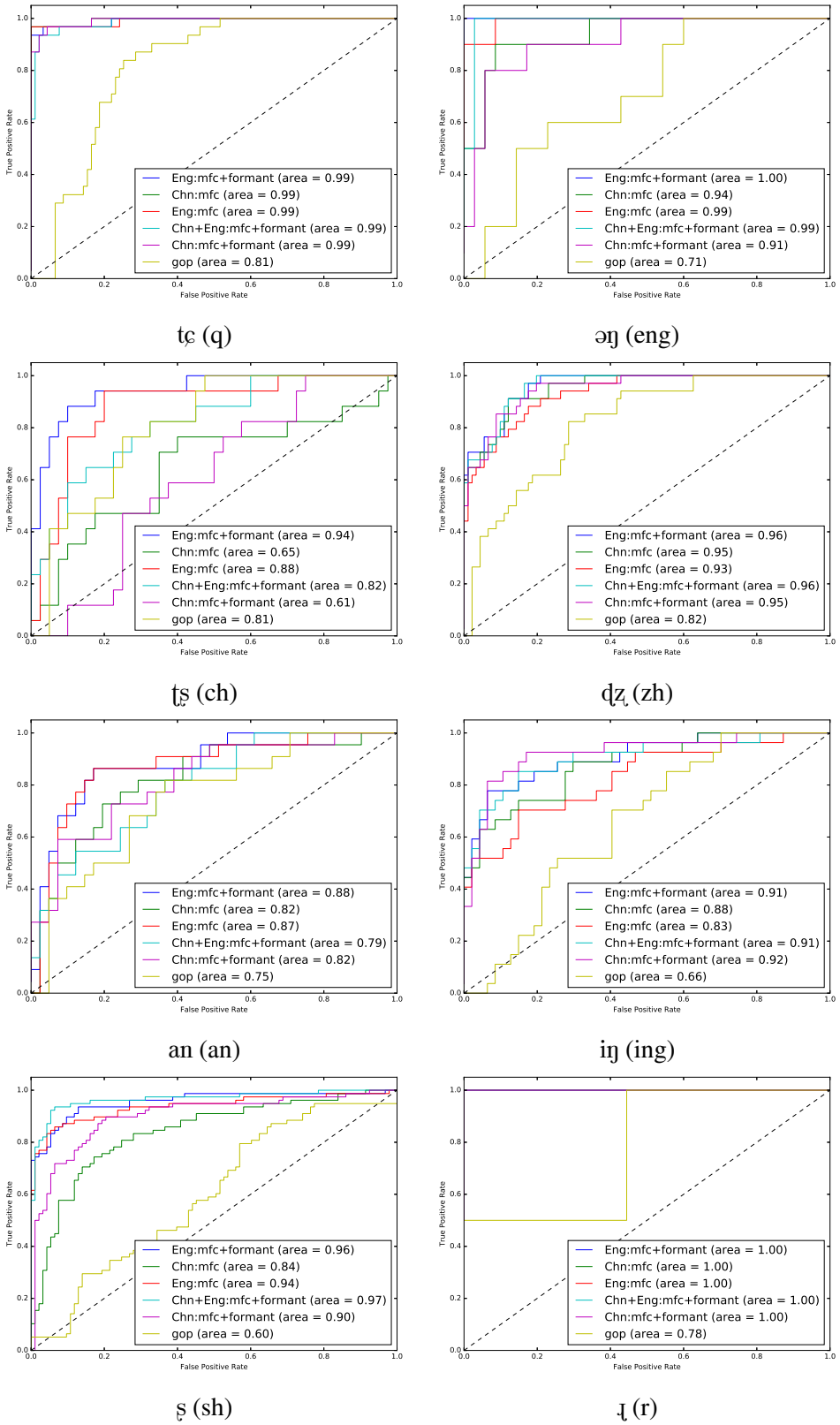


Figure 5.2: (cont.) ROC curves of evaluations on held-out test set for each phone.

positive rate (FPR) can mitigate L2 learner's concerns. Error classes was assigned as positive labels. Figure 5.2 illustrates the curves of comparison results on 16 phones. TPR is on Y-axis, and FPR is on X-axis. This means the top left corner of the plot is the ideal point (TPR=1, FPR=0). Namely, larger area under curve (AUC) indicates better performance. The steepness is also an important sign since the TPR is maximized while keeping FPR minimized.

GOP model (yellow) proved to be a strong baseline for most of the phones, particularly for the phones /uɔ/(uo) and /y/(v) that outperformed all other landmark-based models (AUC>0.9). However, for the phones /ɕ/(x) and /u/(u), the ROC curves of GOP have intercepts with dashed "chance" line, and TPR remains to be zero even when FPR decreases. All other landmark-based models achieved the performance above the "chance" line except for the combined acoustic cues of MFCC and formants at Chinese landmarks on the phone /t͡ʂ/(ch).

5.6 Conclusions

In this chapter, we proposed two approaches to select Mandarin Chinese salient phonetic landmarks for Top-16 frequently mispronounced phonemes by Japanese learners, and extract features at those landmarks including mel-frequency cepstral coefficients (MFCC) and formants. One is to directly map well-founded English landmark theory into Chinese language since there exists correspondences of articulatory-manner and articulatory-place between English and Mandarin Chinese after applying Stevens theory. Second, we defined distinctive Chinese landmarks for Top-16 frequent pronunciation errors by conducting human speech perception experiments in collaboration with linguists.

In order to make fair comparison, we selected a strong baseline model using goodness of pronunciation (GOP). Experiments including 10-fold cross validation on the training set and evaluation on the held-out test set illustrated that acoustic cues of MFCC and formants at both Chinese landmarks and English landmarks led a better performance significantly. When comparing the performance between these two landmark theory, English landmarks locating at both the start and end of phonetic segments for most of the 16 phones slightly outperformed Chinese landmarks that was defined by the empirical analysis of error pairs in the large scale corpus. Chinese landmarks might lose some significant information on discriminating pronunciation errors especially for the nasal phones and fricative

phones. We expected to get access to even larger corpus that are suitable for us to consolidate our Chinese landmark theory.

CHAPTER 6

LANDMARK DETECTION BASED ON CTC AND ITS APPLICATION TO PRONUNCIATION ERROR DETECTION

Acoustic features extracted in the vicinity of landmarks have demonstrated their usefulness for detecting mispronunciation in our recent work [14, 110]. Traditional approaches of detecting acoustic landmarks rely on annotations by linguists with prior knowledge of speech production mechanisms, which are laborious and expensive. This chapter proposes a data-driven approach of connectionist temporal classification (CTC) that can detect landmarks without any human labels while still maintaining consistent performance with knowledge-based models for stop burst landmarks. We designed an acoustic model to predict phone labels based on a recurrent neural network (RNN) with bidirectional long short-term memory (BLSTM) units, which is trained by CTC loss. We found that the positions of spiky phone outputs of this model are consistent with the landmarks annotated in the TIMIT corpus. Both data-driven and knowledge-based landmark models are applied to detect pronunciation errors of second-language Chinese learners. Experiments illustrate that data-driven CTC landmark model is comparable to knowledge-based model in pronunciation error detection. The fusion of them can further improve performance.

6.1 Introduction

Computer-assisted pronunciation training (CAPT) systems provides a flexible and efficient way for the second language (L2) learners to enhance their speaking skills. Pronunciation error detection, as an indispensable part of CAPT systems, has attracted much more attention from different research fields [111].

A number of approaches have been presented to detect pronunciation errors at segmental level in the last few decades. Most of them are based on automatic speech recognition (ASR) frameworks [89,109,112], and they have advantage of predicting pronunciation errors easily and flexibly in same way for all phonemes. However,

they are heavily limited by the size of training data and language backgrounds [113]. And for specific error detection, their detection accuracy need to be improved in order to give learners more precise feedbacks. For L2 learners, the challenge in learning foreign language pronunciation lies in realizing phonetic contrasts since they do not exist in the mother tongue of L2 learners or they do exist but are not phonologically distinctive [96]. For instance, due to the lack of /r/ sound in Japanese, native speakers of Japanese are usually prone to pronounce /r/ as /l/ when they are speaking Chinese. Comprehensive diagnosis of pronunciation errors, such as phonetic minimal pairs, is beneficial for language learners, while the feedback of common ASR errors, i.e. insertions, deletions, or substitutions, only provides limited help. Pronunciation erroneous tendency [109] defined a set of incorrect articulation configurations regarding manners and places of articulation. Several studies have developed classifiers using acoustic-phonetic features in order to capture the subtle distinctions in minimal pairs. Gao et al [114] proposed to detect aspirated stops (/p/, /t/, /k/) in Mandarin Chinese pronounced by native speakers of Japanese on voice onset time features using support vector machines. Strik et al [91] employed the linear discriminant analysis to discriminate a plosive (/k/) from a fricative (/x/) in Dutch using energy-based features with duration information, such as the rate of rise values. The substitution pronunciation errors of Dutch vowels were identified using the first three formants and intensity of segments [96].

It remains difficult to find prominent features for all kinds of phonetic contrasts although aforementioned features took effect in some tasks. Acoustic landmark theory [60] explores underlying regions of quantal nonlinear correlates between articulators and acoustic attributes, and therefore, could provide a cue for choosing distinctive features that are suitable for automatic speech recognition [115] and pronunciation error detection [14, 101, 102, 110]. However, annotating accurate positions of acoustic landmarks is laborious and expensive since it requires solid prior knowledge of speech production mechanism and needs thorough experiments of human speech perception [43, 110]. In order to efficiently obtain accurate annotations of landmarks, we propose an alternative approach by investigating the correlations between acoustic landmarks and locations of spiky phone predictions from connectionist temporal classification (CTC). Recent progress on the end-to-end modeling with CTC [116] has achieved outstanding performance in ASR [117] and keyword spotting [118]. CTC introduced an additional blank symbol to represent the frame-wise phone predictions with very low confidence, so that

RNN-based acoustic models that are trained with the CTC loss ultimately predict pronunciation units with pulse signals, i.e. spikes or peaks. The locations where such spiky predictions occur look similar to the locations of acoustic landmarks in that both underlying acoustic events are assumed to be randomly distributed in the speech signals.

In this chapter, we would verify the consistence between manual annotations of acoustic landmarks and the automatic machine annotations from the data-driven approach using CTC. The landmark annotations from both approaches are applied to the task of pronunciation error detection for Mandarin Chinese learners. Section 6.2 briefly overviews CTC algorithm, the peak detection algorithm, and our framework of pronunciation error detection. Experimental analysis is illustrated in Section 6.3, followed by the conclusion in Section 6.4.

6.2 Methods

6.2.1 Connectionist Temporal Classification

Automatic speech recognition transcribes a sequence of acoustic features into a sequence of words. The sequence of acoustic feature is usually much longer than its corresponding word sequence, therefore, training such models really needs accurate time-alignments of phones in advance. Connectionist Temporal Classification (CTC) relaxes the constraints of needs of prior alignment knowledge, and directly optimizes such sequence to sequence problem by introducing an additional blank symbol which helps to represent predictions of pronunciation units with very low confidence. CTC absorbs ambiguous boundaries between two modeling units and allows repeated labels to appear. Recurrent neural networks (RNNs) are usually applied before CTC calculation, and the softmax output layer computes posteriors probability distributions over all target symbols including phones and the blank at each time step. CTC sums up the probabilities of all possible alignments between labels and input frames by using a forward-backward algorithm. These alignments are implicitly inferred and all lead to the same target label sequence by removing blank symbols and merging repetitions. Figure 6.1 demonstrated the frame-wise phone predictions using CTC.

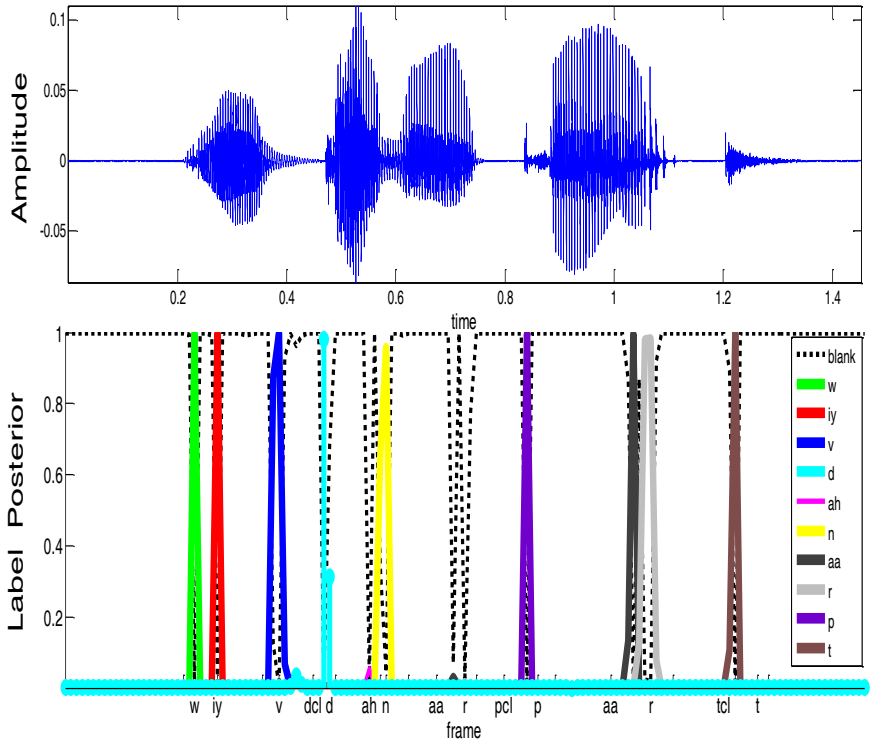


Figure 6.1: Label posteriors estimated by RNN acoustic model trained by CTC on a held-out utterance from "we've done our part".

6.2.2 Peak Detection Algorithm

As shown in Figure 6.1, the BLSTM-RNN acoustic model trained by CTC technique can estimate spiky label posteriors separated by blank labels. CTC utilizes one frame (key frame) with highest posteriors to represent the modeling unit at last. This characteristic is similar to landmark. We assume that the positions of key frames are landmarks. To verify this hypothesis, we should consider to locating the positions of these key frames in an utterance. Palshikar [119] proposed a peak detection algorithm in time-series. We extend it to this work. An utterance is selected as a processing unit. Steps in the algorithm are as follows.

1. Decode each utterance using BLSTM-RNN acoustic model.
2. Extract posteriors of phones detected at each time step. Consequently, it forms a one-dimensional sequence sorted by time index.
3. Compute the peak function value a_i of each point x_i at each time steps. Select the function of $s(\cdot)$ as peak function which computes the average of the maximum distances between k left neighbors and right neighbors of x_i. The left neighbors k is set to 4 time steps—a half of phone duration

estimated on the corpus.

$$s\left(k,i,x_i,T\right)=\frac{1}{2}\left(\max_{1\leq m\leq k}\{x_i-x_{i-m}\}+\max_{1\leq n\leq k}\{x_i-x_{i+n}\}\right)$$

4. Compute the mean and standard deviation of all positive values of a_i.
5. Remove small local peaks in global context according to Chebyshev Inequality and store their temporal information.
6. Order peaks again by their temporal index.
7. Post-processing is to remove adjacent pair of peaks within k time steps.

6.2.3 CTC-Based Landmark Detection

We evaluated the performance of CTC-based landmark detection on TIMIT English corpus because it defined a mature labeling convention of landmarks and provided accurate landmark annotations as well. Miller et al [120] shows that it reaches highest annotation accuracy about stops among the manually annotated phones. Stop burst is realized by a sequence of two coherent events: a closure and a release, and it provides a salient distinctive cue to identify stop consonant. We implemented a ASR system based on CTC, and a greedy search method was applied to predict phone labels from their posterior distribution in each utterance. And then aforementioned peak detection algorithm was implemented to locate the candidate landmarks.

6.2.4 Pronunciation Error Detection

Figure 6.2 illustrates the overall framework of pronunciation error detection. It is divided into two stages. In the first stage, the BLSTM-RNN acoustic model is trained with a large native corpus. And the acoustic model estimates posteriors of phones which are used as the canonical sounds to estimate landmarks. A peak detection algorithm is used to determine the positions of spiky posteriors of phones in an utterance. Further, to get landmark of one phone, a transcription with time alignments for each phone is used to determine the offset to the onset of this phone. We used the relative positions of landmarks to compare with our previous work [14]. Each relative position is defined as the value of offset divided by the duration of that phone. Finally, the mean value of all landmarks for each kind of phone is calculated.

In the second stage, we construct classifiers for each binary phonetic contrast—mispronunciation and its canonical sound—by extracting acoustic features in the vicinity of landmarks determined at the first stage.

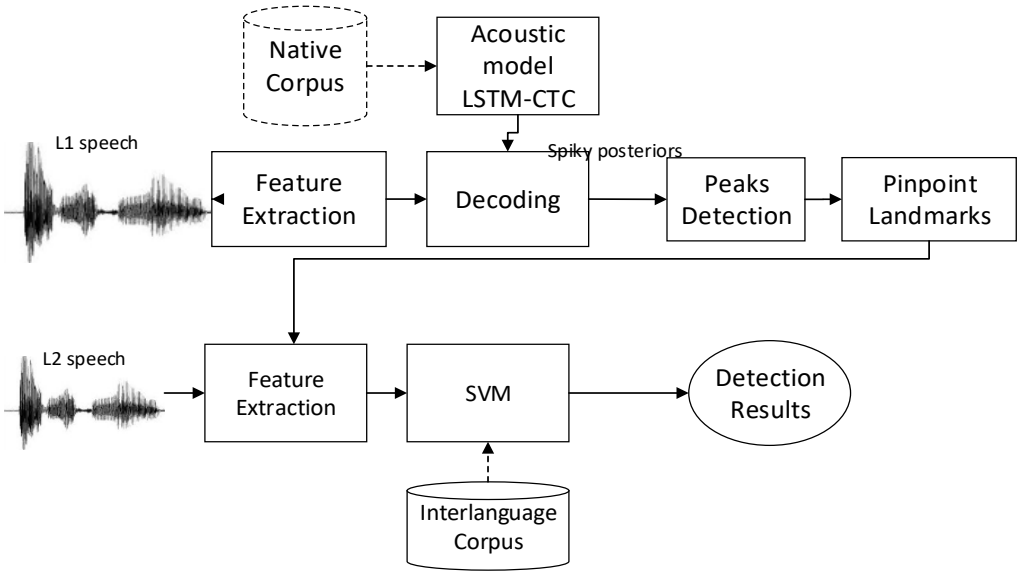


Figure 6.2: The overall framework of pronunciation error detection.

6.3 Experiments and Results

6.3.1 CTC-Based Landmark Detection

Corpora

We first developed a CTC-based ASR system for English phone recognition. We selected the portion of 100 hours of LibriSpeech corpus [10] which consisted of 251 speakers including 126 males and 125 females. There were 28539 labeled utterances. 4000 utterances were selected as the validation set and the rest for training. The TIMIT [121] corpus was selected as the test set, which contained 5040 utterance (except the dialect utterances SA) and 630 speakers (438 males and 192 females). Since landmark annotations in the TIMIT transcriptions were based on abrupt acoustic changes, if no acoustic evidence existed for a certain phone, then no label was put there. About 68% of the total number of landmarks in the corpus are acoustically abrupt landmarks which are associated with consonantal segments,

e.g., a stop closure or release. Therefore, we chose stops (/p/,/t/, /k/,/b/,/d/,/g/) to verify our hypothesis. Their landmarks are at the beginning of release of stops (marked with /pcl/, /tcl/, /kcl/, /bcl/, /dcl/, /gcl/) or at the end of closure. We extracted 31256 stop burst landmarks from the transcriptions.

Experimental Setup

Experiments were conducted for phone recognition using EESEN [122]. The BLSTM-RNN acoustic model was trained by CTC loss. The fliterbank features (40-dimention) with their first and second order derivatives were extracted. Each speech frame span across 25ms window and it shifted by every 10ms to the input layer of BLSTM-RNN. Speaker normalization was applied to the speech features. Four BLSTM layers contained 320 cells in each direction. We employed the CMU dictionary[1] as the lexicon while ignoring the stresses. We extracted 43 labels from CMU dictionary including phones, noise marks, and the additional blank. The weights of the model were initialized uniformly from the range $[-0.1, 0.1]$. The model was trained by back-propagation through time with the initial learning rate 0.0004.

Evaluation Metrics

Three evaluation metrics are used to compare consistency [123]. If a detected landmark falls in the region of hand-labeled landmarks with a certain time tolerance, it is considered as a hit.

- Recall: The ratio of the number of hits to the number of hand-labeled landmarks.
- Precision: The ratio of the number of hits to the number of total landmarks detected.
- F-measure: The harmonic mean of recall and precision.

Results

We conducted different time tolerances from 10ms to 50ms. Performance of detection process for six stops is listed in Table 6.1.

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Table 6.1: Detection results (%) of four stop burst landmarks

Evaluation Metrics	Time Tolerance (ms)				
	10	20	30	40	50
Recall	74.2	81.4	83.8	85.6	87.2
Precision	61.8	67.8	69.8	71.3	72.6
F1 score	67.4	74.0	76.2	77.8	79.2

As shown in Table 6.1, we can see that CTC-based method can detect most of landmarks in certain time tolerance for six stops. The slightly low precision values indicate relatively high insertion errors. According to our statistics, there are many cases such as co-articulation and allophones in the running speech where there are no labels or labeled by allophones. For example, there is "but didn't" in the transcription `train/dr2/fajw0/sx273.phn`. /t/ is in the final position of 'but' and /d/ is in the initial position of "didn't". There is no release of /t/ and there is no closure of /d/. For allophones, sometimes /t/ can also be realized as a glottal stop annotated as /q/, such as in "cotton". A flap labeled with /dx/ can either be an underlying /t/ or /d/. Many of /t/ and /d/ may not manifest obvious obstruction in the running speech, therefore, no corresponding landmarks are annotated. It is very difficult to locate the landmarks in the running speech, so that CTC-based method almost generates a peak for each phone, resulting in increases of insertion errors. But in any case, this method may help linguists annotating landmarks in some complicated cases, and it could be a good tool of front-end processing for pronunciation error detection.

6.3.2 Landmark-Based Pronunciation Error Detection

Corpora

The Chinese National Hi-Tech Project 863 corpus [124] of approximate 100 hours was used to train acoustic model, which consisted of 82735 utterances. 4000 utterances were selected from the training data as the validation set. 6 native Chinese speakers from Chinese part of BLCU inter-Chinese speech corpus [106] were used for estimating landmarks. Their utterances were firstly force-aligned by HTK [107] and then the phonetic boundaries were corrected by human transcribers. For the purpose of CAPT, we collected a large scale of Chinese inter-language

corpus read by Japanese native speakers, which is referred to as the BLCU inter-Chinese corpus [106]. We select 7 Japanese females who read 1899 utterances. 80% of the data was selected as the training set and the rest as the test set. 16 most frequent PETs [109, 112] and their canonical sounds constituted 16 binary phonetic contrasts. Their training set and test set were selected from above partition.

Experiment Setup

A CTC-based speech recognition system was built firstly. The architecture of network was identical to the description in the last experiment. There were 60 labels (including Chinese phones, noise marks and blank) in the softmax layer. Stevens proposed that distinctive features obtained from landmark regions should be universal across languages. By applying this theory, Yang et al [14] found the correspondence between English and Mandarin Chinese from the perspective of manner and place of articulations, and English phones were mapped to Chinese phones with the guidance of international phonetic alphabet (IPA). Both knowledge-based landmarks and our data-driven landmarks are summarized in Table 6.2.

Table 6.2: Data-driven and knowledge-based landmarks.

Phone	English Landmark	CTC Landmark
ʂ (sh)	start, end	onset (0.307) of consonant
ʤʐ (zh)	start, end	onset (0.271) of consonant
ʧʂ (ch)	start, end	onset (0.217) of consonant
ɕ (x)	start, end	onset (0.356) of consonant
ʤʑ (j)	start, end	onset (0.310) of consonant
an (an)	start,end	onset (0.073) of vowel
y (v)	middle	onset (0.030) of vowel
aŋ (ang)	start, end	onset (0.058) of vowel
iŋ (ing)	start, end	onset (0.055) of vowel
u (u)	middle	onset (0.194) of vowel
f (f)	start, end	onset (0.401) of consonant
əŋ (eng)	start, end	onset (0.079) of vowel
tɕ (q)	start, end	onset (0.284) of consonant
k (k)	start	onset (0.253) of consonant
ɻ (r)	start, end	onset (0.392) of consonant
uɔ (uo)	middle	onset (0.342) of vowel

According to this table, three frames from the landmarks were selected. We used 13-dimentional Mel-frequency cepstral coefficient (MFCC) features with their first

and second derivatives to obtain a total of 39 feature values per frame. To capture subtle variations and to address the issues of insufficient number of frames, 6ms frame shift was selected. The MFCC features of 3 frames from landmarks were concatenated to form a 117-dimensional vector. We used SVMs with linear kernel as a classifier.

Detection Results

We compared two kinds of landmarks, data-driven and knowledge-based, and combined them at last. We used F-measure to evaluate their performance. Table 6.3 shows that CTC landmark system is better or comparable to knowledge-based landmark system when detecting seven phones (/an/, /f/, /ing/, /j/, /k/, /u/, /v/). The performance of the former is reduced more than 5% in comparison to the latter when detecting five phones (/ch/, /q/, /r/, /sh/, /zh/). This is because different positions and amounts of landmarks may affect the results. The fusion of them further improves the performance (except /an/).

Table 6.3: Performance of landmark-based pronunciation error detection.

Phone	English Landmark	CTC Landmark	Fusion
ʂ (sh)	0.784	0.894	0.883
ɖʐ (zh)	0.844	0.940	0.928
ʈʂ (ch)	0.817	0.876	0.886
ɕ (x)	0.940	0.975	0.958
dʑ (j)	0.982	0.982	0.982
an (an)	0.864	0.864	0.851
y (v)	0.914	0.893	0.903
aŋ (ang)	0.956	0.981	0.981
iŋ (ing)	0.824	0.820	0.880
u (u)	0.929	0.929	0.929
f (f)	0.935	0.919	0.946
əŋ (eng)	0.932	0.959	0.959
tɕ (q)	0.854	0.973	0.968
k (k)	0.905	0.905	0.905
ɻ (r)	0.889	0.941	1.000
uɔ (uo)	0.964	0.967	0.969

We also compared our system with the hybrid DNN/HMM in our previous work [109] by considering overall detection results. According to [109], three

kinds of metrics were used to evaluate the detection performance as shown in the following:

- False Rejection Rate (FRR): The percentage of correctly pronounced phones that are erroneously rejected as mispronounced.
- False Acceptance Rate (FAR): The percentage of mispronounced phones that are erroneously accepted as correct.
- Detection Accuracy (DA): The percentage of detected phones that are corrected recognized.

Table 6.4 shows that CTC-based landmark system exhibits slightly poor performance compared with knowledge-based landmark system and hybrid DNN/HMM. Knowledge-based landmark system performs better than hybrid DNN/HMM. Two kinds of landmarks are combined further to improve the overall detection performance.

Table 6.4: The results of landmark-based system and DNN/HMM hybridsystem.

System	FRR	FAR	DA
CTC+Landmark+SVM	14.0	30.4	84.4
English+Landmark+SVM	9.1	16.1	89.9
Landmarks Combination+SVM	9.0	16.1	90.0
DNN/HMM+MFCC	6.7	35.9	87.6

6.4 Conclusions

In this chapter, we firstly verify the hypothesis that the positions of spiky phone posterior outputs of the model trained by CTC loss are consistent with the stop burst landmarks annotated in the TIMIT corpus. As a result, we think these peaks evaluated by CTC-based acoustic model are similar to landmarks and they can be generalized to other phones. Then we propose a pronunciation error detection framework on Chinese learning based on landmarks and SVMs, and the landmarks can be predicted automatically from BLSTM-RNN acoustic model. CTC-based data-driven method and knowledge-based method are all considered to locate landmarks. Experiments illustrate that data-driven CTC landmark model is comparable to knowledge-based model in pronunciation error detection. Their combination

further improves the performance which outperforms DNN/HMM hybrid system with MFCC features.

CHAPTER 7

CONSONANT VOICING DETECTION ON MULTI-LINGUAL CORPORA

This chapter tests the hypothesis that distinctive feature classifiers anchored at phonetic landmarks can be transferred cross-lingually without loss of accuracy. Three consonant voicing classifiers were developed: (1) manually selected acoustic features anchored at a phonetic landmark, (2) MFCCs (either averaged across the segment or anchored at the landmark), and (3) acoustic features computed using a convolutional neural network (CNN). All detectors are trained on English data (TIMIT), and tested on English, Turkish, and Spanish (performance measured using F1 and accuracy). Experiments demonstrate that manual features outperform all MFCC classifiers, while CNN features outperform both. MFCC-based classifiers suffer an overall error rate increase of up to 96.1% when generalized from English to other languages. Manual features suffer only an up to 35.2% relative error rate increase, and CNN features actually perform the best on Turkish and Spanish, demonstrating that features capable of representing long-term spectral dynamics (CNN and landmark-based features) are able to generalize cross-lingually with little or no loss of accuracy.

7.1 Introduction

In contrast to the conventional data-driven speech recognition model, acoustic correlates of distinctive features are found in an acoustics phonetic recognizer [31] so as to extract interpretable acoustic information. There are two types of distinctive features in this model: articulator-free feature and articulator-bound feature. Articulator-free features determine phone manner class, while articulator-bound features specify the phone identity. Since features in this model depend on the properties of the vocal tract, they are, to some extent, universal and independent of the language being spoken. For obstruent consonants in English, such as fricatives, affricates, and stops, three articulators ([lips], [tongue body] and [tongue blade])

can form the constriction to produce consonants. Obstruent consonants are further categorized by consonant voicing which can be described by the articulator-bound feature [stiff vocal folds] [125].

Much related work has been done about consonant acoustic and voicing. Shadle [126] studied fricative consonants using mechanical models, theoretical models, and acoustic analysis, and found that the most important parameters for fricatives are the length of the front cavity, the presence of an obstacle and the flow rate. Speech production mechanism differences between voice and voiceless stops are mainly due to muscle activity, which relaxes the tongue root during voiced stops, altering aerodynamics near the vocal folds in order to maintain voicing during closure [127]. Vowel cues, such as vocalic duration and F1 offset frequency, are also the correlates of consonant voicing [128]. A module for detecting consonant voicing based on these acoustic correlates [46] first determines acoustic properties according to consonant production, then extracts acoustic cues, and classifies them to detect consonant voicing.

One of the traditional methods to detect consonant voicing uses MFCCs [129, 130]—voicing can be detected with 74.7% to 80% overall accuracy [53, 131]. Achieving good performance of MFCC-based method on consonant voicing is possible, however, MFCCs are less efficient in capturing information about the voice source because MFCCs mostly codify the "filter" information in the source-filter theory of speech production.

In contrast, much of the consonant voicing information can be captured in the characteristics of the vocal fold vibration patterns, therefore capturing acoustic phonetic features indicative of vocal fold vibration has the potential to measure consonant voicing. Though voicing does not continue uninterrupted during obstruent closure in English, there are striking differences near consonant closure and release landmarks. Landmarks [31] identify times when the acoustic patterns of the linguistically motivated distinctive features are most salient; acoustic cues extracted in the vicinity of landmarks may therefore be more informative for the classification of distinctive features than cues extracted from other times in the signal. To the best of our knowledge, the highest accuracy for voicing classification of obstruents uses acoustic features extracted with reference to phonetic landmarks, with accuracies of 95% and 96% [132, 133] for stops and fricatives respectively.

The choice of data representation is essential for the performance of detection or classification tasks. Discriminative information from raw data can be extracted by taking advantage of human perceptual ingenuity and human prior knowledge.

However, the process of designing these manual features is laborious and time-consuming. Deep learning techniques transform raw data into multiple levels of abstraction by stacking multiple layers with non-linearities, thus learning complex features automatically [134]. Though the accuracy of speech recognizers built from deep networks is high [135], results on the cross-language portability of deep networks include both positive and negative outcomes. We propose that deep networks trained to classify distinctive features should be cross-language portable, because of the universality of the features they are trained to classify.

In order to test the hypothesis that distinctive features anchored at phonetic landmarks can be transferred cross-lingually, we train models on an English corpus and evaluate them on voicing detection tasks for three different languages including English, Spanish, and Turkish.

In the following sections, acoustic landmark theory and the definition of its regions are described in Section 7.2. Within these landmark regions, Section 7.3 illustrates acoustic feature representations that help to improve the performance of voicing detection, consisting of manually designed acoustic cues and features learned from deep neural networks. Experiments and analysis are described in Section 7.4 and Section 7.5, followed by conclusions in Section 7.6.

7.2 Acoustic Landmarks and Distinctive Features

Landmarks are defined as points in an utterance where information about the underlying distinctive features may be extracted. Four types of landmarks were proposed in [31]: vowel (V), consonant release (Cr), consonant closure (Cc), and glide (G). Cr and Cc landmarks are further specialized by manner classes—stop (S), fricative (F), and nasal (N). For example, nasal release and closure can be defined as Nr and Nc respectively.

We assume that accurate landmark positions in a speech signal are provided, so that TIMIT phonetic transcriptions can be converted into landmark transcriptions under the following rules. Each stop release segment has a Sr landmark at its start time; each stop closure segment has a Sc landmark at its start time; each affricate, fricative, or nasal has a Cc landmark at its start time and a Cr landmark at its end time, specifically, each affricate and fricatives has a Fc and Fr, and each nasal has a Nc and Nr; each vowel and glide has a landmark located at the midpoint of its duration. TIMIT transcriptions specify the time-alignments of each phone

with its acoustic signal, therefore, computing and generating landmark labels is pretty straightforward by applying the aforementioned rules. Table 7.1 illustrates examples of acoustic landmarks extracted from TIMIT. The first column denotes the time when landmarks exist; the second column displays detailed landmark types accordingly. For example, a fricative closure (Fc) landmark happened at the time 0.1916s as shown in the first row in the table.

Table 7.1: Examples of landmark transcription from TIMIT.

Time (s)	Landmark type
0.1916	Fc
0.2839	Fr
0.3213	V
0.3864	G

Stevens [31] proposed that distinctive features obtained from closure and release landmark regions should be universal across languages. Motivated by this theory, landmark positions across multiple languages can be labeled by the same rules. In this chapter, we consider corpora in three languages—English, Spanish, and Turkish. Landmark regions are further defined in the following way once landmark positions are pinpointed.

- If a Cc landmark happens at the start of a phone, acoustic cues are extracted at the time when 20ms is delayed from the start (+20ms).
- If a Cr landmark happens at the start of a phone, acoustic cues are extracted at the time when 20ms is advanced from the start (-20ms).

Acoustic cues correlate with manners of articulation so that they determine the activation of distinctive features. Distinctive features are the concise description of subsegmental attributes of a phone with a relatively direct relationship to acoustics and articulation. These features typically take on binary values and form a minimal set which can help to distinguish each segment from all others in a language. The phonetic transcription of an utterance is thereby obtained if a collection of categorical distinctive features can be detected. In this chapter, we are interested in detecting consonant voicing, one of distinctive features.

7.3 Acoustic Feature Representations

This section would describe the extraction of expressive acoustic features that help to distinguish voiced consonants from unvoiced ones. We would elaborate hand-crafted acoustic cues as well as self-learned features by deep neural networks.

7.3.1 Hand-Crafted Acoustic Cues

We explored several acoustic features using conventional speech signal processing techniques, and a summary of these features is shown in Table 7.2.

Voice onset time (VOT): The duration between a consonant release and the onset of voicing is demonstrated to carry voicing information for English stops, fricatives, and affricates [131]. The voiced consonants are usually signaled in a shorter duration than unvoiced ones.

Peak of normalized cross-correlation (PNCC): The value of cross-correlation will be increased whenever the voicing of stops, fricatives, and affricates is produced [31]. For the voiced speech, the glottal period usually varies by only a small percentage from one period to the next; the vocal tract filter varies slowly in comparison to the glottal inter-pulse interval so that adjacent periods of the speech signal tend to have similar shapes. Talkin [136] discovered a generator of candidate estimates for the true period of speech signals based on the normalized cross-correlation function (NCCF). We retain the peak of NCCF to capture the value transitions of cross-correlation.

Amplitude of fundamental frequency (H1): Amplitudes of the speech signal varies in time, however, observations between voiced and unvoiced segments show that the amplitude of unvoiced speech is usually smaller than voiced speech. We extract the amplitude of fundamental frequency as another feature in order to account for the strength of vocal fold vibration.

Formant transitions: The behaviors of formant transitions are different between voiced and unvoiced consonants [137]. We can observe an obvious formant transition appearing right after the voice onset in voiced obstruents, while on the contrary, there is not obvious similar behavior happened for unvoiced obstruents. The transition behavior is significant for stops, and is also observable for other obstruents.

Energy: The energy distribution over frequency bands deviates between voiced and unvoiced consonants. It is observable for voiced consonants that most energies

spread on low and relavtive high frequencies, while most energies are concentrated at high frequencies for unvoiced consonants. Movitivated by the observation, we extract energy-based features including root-mean-square (RMS) energy, energy between 0-400Hz (E1), energy between 2000-7000Hz (E2), and the ratio of E1 and E2.

Table 7.2: Hand-crafted acoustic features in landmark regions.

Acoustic Feature	Motivation
RMS energy Energy between 0-400Hz (E1) Energy between 2000-7000 Hz (E2) Ratio of E1 and E2	Energy distributions over frequency bands are different between voiced and unvoiced consonants
Peak of normalized cross-correlation (PNCC)	PNCC increases when vocal folds are vibrating
Amplitude of fundamental frequency (H1)	Amplitude of voiced speech is larger than unvoiced speech
Voice onset time (VOT)	Duration between release and voicing onset is shorter for voiced segments than unvoiced segements
Formants transition	Voiced obstruents have obvious formant transitions while unvoiced ones do not have

7.3.2 Self-Learned Features by Convolutional Neural Networks

Common feature representation of speech signals as inputs to deep neural networks is the spectrograms—magnitudes of log-Mel filter banks over time. However, this feature may not be suitable for our task. We are looking into the features extracted within a landmark region that is typically too short to contain multiple speech frames. Therefore, we instead only perform the convolution operators along frequency band.

Figure 7.1 illustrates the architecture of convolutional neural networks (CNN) that consist of three types of layers—convolutional, max-pooling, and fully connected layers. In a convolutional layer, each neuron takes as inputs local patterns in the previous layer. All neurons in the same feature map share the same weight matrix. A max-pooling layer is stacked following each convolutional layer that similarly takes local patterns as inputs, and down-samples to generate a single output for that local region. Multiple fully connected layers are concatenated after multiple building blocks of convolutional-pooling pairs. A softmax layer with a single neuron is taken as the output that capture the posterior probability of the positive label (consonant voicing). During back-propagation, a first-order gradient-based optimization method based on adaptive estimates of lower-order moments (Adam) [138] is used.

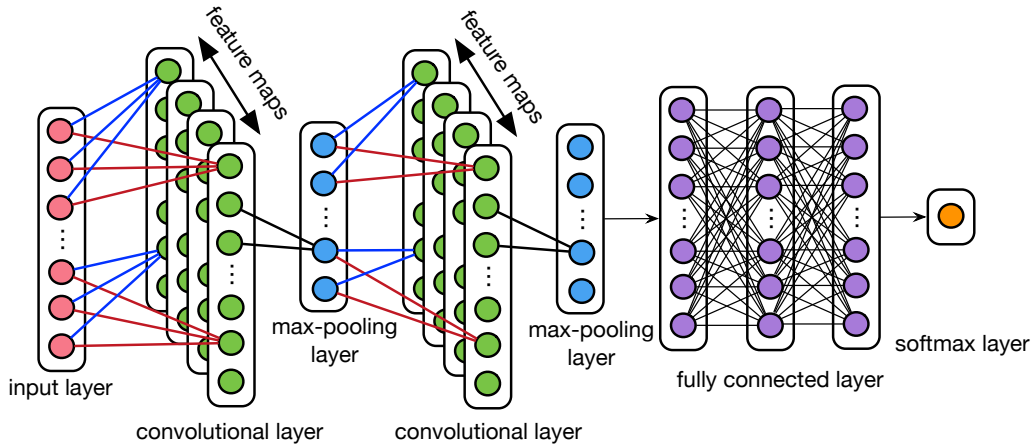


Figure 7.1: Convolutional neural networks for speech signals

7.4 Experiments

7.4.1 Multilingual Corpora

We consider three languages including American English, Spanish, and Turkish. American English corpus is used for model training, and all other languages are used for model evaluation. Table 7.3 shows the basic statistics of corpora for three languages.

Table 7.3: Data statistics for consonant voicing detection across three languages.

	Voiced consonant	Unvoiced consonant
TIMIT (train)	56,269	40,475
TIMIT (test)	20,769	14,214
Spanish (test)	68,946	24,529
Turkish (test)	13,179	4,722

American English Corpus: TIMIT [56] corpus contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences and includes time-aligned orthographic, phonetic, and word transcriptions. Each utterance is recorded as a 16-bit and 16kHz speech waveform file.

Spanish corpus: The phonetic Albayzin corpus of Spanish, initially developed to train speech recognition engines, is selected. It was further divided into training and testing subsets. The training data is made of 200 phrases; 4 speakers produced

all 200 phrases and 160 speakers produced 25 out of these 200, so the set of 200 phrases is produced 24 times. The phrases are acoustically balanced according to a statistical study of the frequency of each sound in Castillian Spanish.

Turkish corpus: Middle East Technical University Turkish Microphone Speech Corpus (METU) [139] was selected as Turkish test set. 120 speakers (60 male and 60 female) speak 40 sentences each (approximately 300 words per speaker), which makes around 500 minutes of speech in total. The 40 sentences are selected randomly for each speaker from a triphone-balanced set of 2462 Turkish sentences.

7.4.2 Feature Extraction

This section illustrates details of calculation of acoustic features including hand-crafted acoustic features anchored at phonetic landmark regions, MFCCs that are either averaged across the phonetic segment or anchored at the landmark region, and self-learned CNN features in landmark regions.

MFCCs: A Hamming window is applied on the duration of the landmark region or of the whole phone. The windowed signal is then transformed to compute Mel-frequency cepstral coefficients (MFCC). We calculate static coefficients (MC13) and their combination with dynamic coefficients (MC39).

VOT, formant transition, PNCC and H1: A robust RAPT [136] algorithm for pitch tracking that is based on normalized cross-correlation and dynamic programming is applied using Wavesurfer[1]. The fundamental frequency, probability of voicing (1.0 means voiced and 0.0 means unvoiced), local error of the pitch, and the peak of normalized cross-correlation are obtained. After getting the pitch for each landmark segment, FFT amplitude at the pitch frequency was measured, and FFT spectra were used to measure formant transitions.

Energy: Butterworth filter is used to design a bandpass filter with ≤ 3dB of passband ripple and ≥ 40dB attenuation in the stopbands. Energies of filtered signals are then computed in the frequency bands between 0-400Hz and between 2000-7000Hz, respectively.

Self-learned CNN features: 1024 point magnitude FFT is performed. Mel-scale filterbank features are calculated by multiplying frequency response with a set of 40 triangular bandpass filters equally spaced in Mel frequency. In order to apply the early stopping strategy during training procedure, a held-out development set

[1] `http://www.speech.kth.se/wavesurfer/`

(10%) is stratified sampled from training set. The training will stop when the validation loss is not decreasing anymore within 10 consecutive epochs.

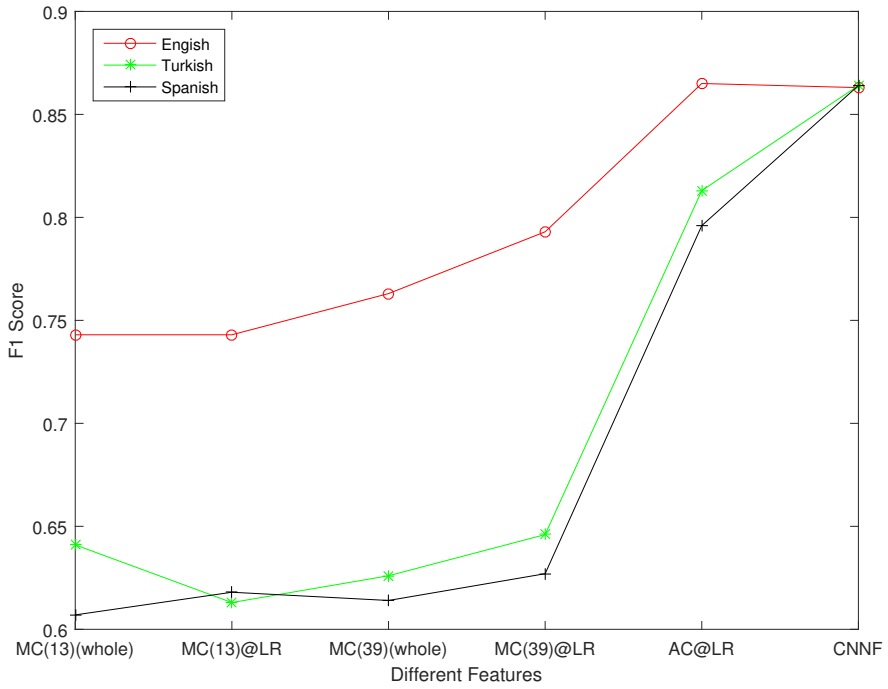
7.5 Results

Consonant voicing is detected using MFCCs, hand-crafted acoustic cues, and self-learned CNN features. These models are all trained in English, and tested on English, Spanish, and Turkish, respectively. Support vector machine with radial basis function kernel is used as the binary classifier based on acoustic features, while CNNs are used as an end-to-end classifier. The F1 score of voicing consonant (positive sample) and overall accuracy are used as metrics. Due to the imbalanced nature of training set, the CNN is trained with each sample weighted inversely proportional to class frequency. Relative error rate increment of performance over English has been calculated when models are applied on other languages.

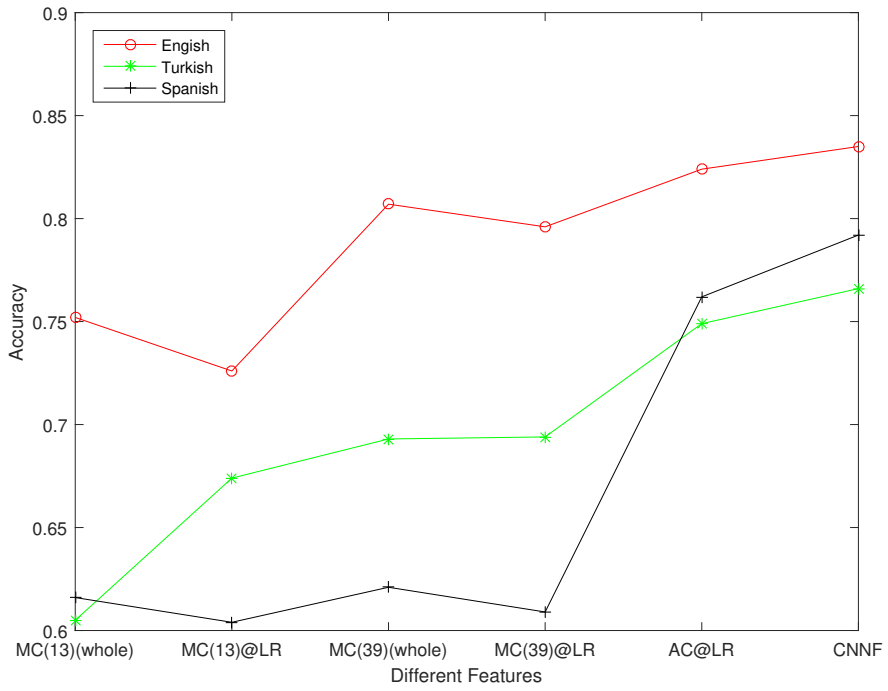
MFCCs: When calculating average MFCCs across the whole phonetic segment, MFCCs with dynamic coefficients (MC39) achieved better accuracy and F1 score than MFCCs with only static coefficients (MC13) as shown in the first and third columns of Figure 7.2(a) and 7.2(b). However, F1 and accuracy for MFCCs dropped by 5-20% absolutely when these models were evaluated on Spanish and Turkish. When calculating MFCCs anchored at landmark regions, MC39 obtained slightly better F1 score than its averaged model across the whole segment.

Acoustic Cues vs MFCCs: Acoustic cues demonstrated the superb performance over MFCCs on all three languages as shown in Figures 7.2(a) and 7.2(b). Models with acoustic cues trained on English suffered much smaller reductions of accuracy and F1 score than MFCCs based models when tested on the other two languages: MFCCs models achieved up to 96.1% accuracy relative reduction while acoustic cues model only achieved up to 35.2%.

CNNs vs Feedforward NNs: The filterbank used to compute MFCCs can be viewed as a type of pre-determined convolutional network; conversely, CNNs extract local patterns with trainable but fixed-length convolutional windows. The last two columns in Figure 7.3(a) reveal that CNNs can hold stable performance for each language, using either FFT or filterbank features as inputs. When applied to Spanish and Turkish, CNNs show little drop in accuracy, while their F1 score is higher in the test languages than in the training language as shown in Figure 7.3(b); the difference between overall accuracy and F1 score is apparently an artifact of

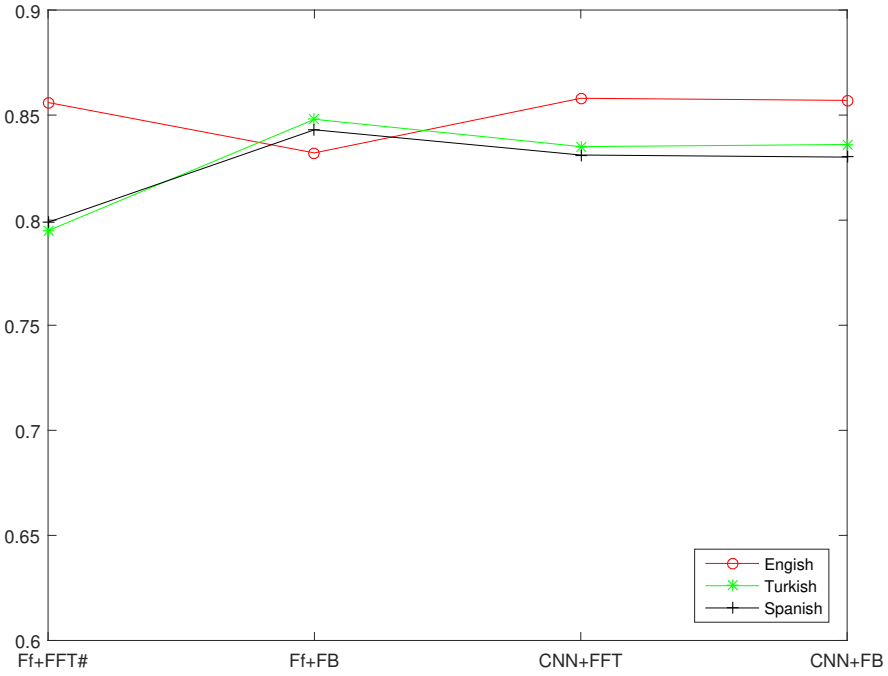


(a) F1 score

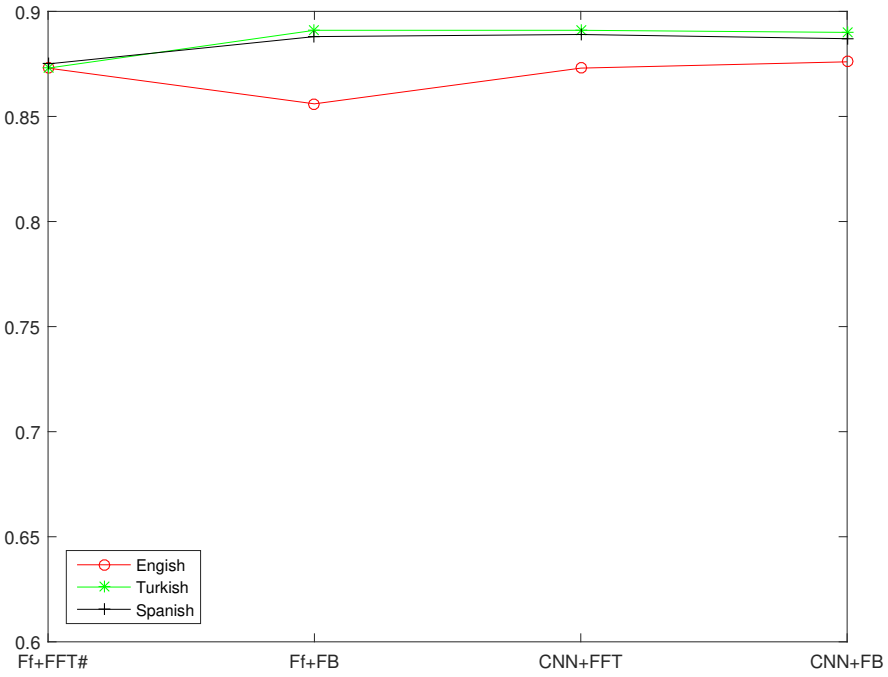


(b) Accuracy

Figure 7.2: Performance for three different features ("MC", "AC", and "CNN") on three test languages (English, Spanish, and Turkish), where "MC" denotes MFCC, "AC" denotes acoustic cues, "CNN" denotes acoustic features extracted by CNN, "LR" denotes landmark region, "whole" denotes whole duration of a phone.



(a) Accuracy



(b) F1 score

Figure 7.3: Performance for different neural network configurations, where "Ff" denotes a feedforward neural network and "FB" denotes filterbank features.

the highly non-uniform class distribution in Turkish and Spanish, both of which have twice as many voiced as unvoiced obstruents (see Table 7.3).

CNNs vs Acoustic Cues: Since the evaluation on two models—CNN+FFT and CNN+FB—results in similar accuracy scores, we consider CNN+FFT as the best CNN model. The right most points in Figures 7.2(a) and 7.2(b), and the number in the last row in Table 7.4 illustrate that the best CNN model outperforms acoustic cues across three languages, in both metrics, and that it generalizes well to cross-lingual corpora.

Table 7.4: Relative accuracy reductions (%) over English for all other languages.

	Turkish	Spanish
MFCC(13) (whole utterance)	59.3	54.8
MFCC(13) (Landmark Region)	40.9	44.5
MFCC(39) (whole utterance)	59.1	96.4
MFCC(39) (Landmark Region)	50.0	91.7
Acoustic cues	31.3	35.2
CNN	16.2	19.0

7.6 Conclusion

In this work, three different features are applied to build consonant voicing detectors, in order to test the theory that distinctive feature-based classes are robust over multilingual corpora. MFCCs (in landmark region, and averaged over the whole phone utterance duration), acoustic features extracted from the landmark region, and features learned by a convolutional neural network (CNN) were tested as features. Classifiers based on these features are all trained on English and tested on English, Spanish, and Turkish. Results show that MFCCs could not capture voicing in either the training language or test languages. Manual acoustic features generalize better to novel languages than MFCC. Acoustic features learned by a CNN obtain best performance, both on training languages and non-training languages. We conclude that features capable of representing long-term spectral dynamics relative to a phonetic landmark (CNN and landmark-based features) are able to generalize cross-lingually with little or no loss of accuracy.

CHAPTER 8

JOINT MODELING OF ACOUSTICS AND ACCENTS

The performance of automatic speech recognition systems degrades with increasing mismatch between the training and testing scenarios. Differences in speaker accents are a significant source of such mismatch. The traditional approach to deal with multiple accents involves pooling data from several accents during training and building a single model in multi-task fashion, where tasks correspond to individual accents. In this section, we explore an alternate model where we jointly learn an accent classifier and a multi-task acoustic model. Experiments on the American English Wall Street Journal and British English Cambridge corpora demonstrate that our joint model outperforms the strong multi-task acoustic model baseline. We obtain a 6.86% relative improvement in word error rate on British English, and 11.34% relative improvement on American English. This illustrates that jointly modeling with accent information improves acoustic model performance.

8.1 Introduction

Recent breakthroughs in automatic speech recognition (ASR) have resulted in a word error rate (WER) on par with human transcribers [2, 3] on the English Switchboard benchmark. However, dealing with acoustic condition mismatch between the training and testing data is a significant challenge that remains unsolved. It is well-known that the performance of ASR systems degrades significantly when presented with speech from speakers with different accents, dialects and speaking styles than those encountered during system training [8]. We specifically focus on acoustic modeling for multi-accent ASR in this Chapter.

Dialects are defined as variations within a language that differ in geographical regions and social groups, which can be distinguished by traits of phonology, grammar, and vocabulary [140]. Specifically, dialects may be associated with the residence, ethnicity, social class, and native language of speakers. For example,

in British and American English, same words can have different spellings, like *favour* and *favor*; or different pronunciations, such as *ˈʃɛdjuːl* in UK English vs. *ˈskɛdʒʊl* in US English for the word *schedule*; in Spanish, vocabulary may evolve differently between dialects, like for the phrase *cell phone*, Castilian Spanish uses *móvil* while Latin American use *celular* [141]; in English, same phoneme may be realized differently, phoneme */e/* in *dress* is pronounced as */ɛ/* in England and */e/* in Wales; in Arabic, dialects may also differ in intonation and rhythm cues [142]. In this chapter, we focus on the issue of differing pronunciations, while eschewing considerations of grammatical and vocabulary differences.

Acoustic modeling across multiple accents has been explored for many years, and various approaches can be summarized into three categories - *Unified models*, *Adaptive models*, and *Ensemble models*. A unified model is trained on a limited number of accents, and can be generalized to any accent [143, 144]. An adaptive model fine-tunes the unified model on accent-specific data assuming that the accent is known [145–147]. An ensemble model aggregates all accent-specific recognizers, and produces an optimal model by selection or combination for recognition [141, 148, 149]. Experiments have revealed that the unified model usually underperforms the adaptive model, which in turn underperforms the ensemble model [143, 144].

We note that these prior approaches do not explicitly include accent information during training, but do so only indirectly, for example, through the different target phoneme sets for various accents. This contrasts sharply with the way in which humans memorize the phonological and phonetic forms of accented speech: "mental representations of phonological forms are extremely detailed," and include "traces of individual voices or types of voices" [150]. In this chapter, we propose to link the training of ASR acoustic models and accent identification models, in a manner similar to the linking of these two learning processes in human speech perception. We show that this joint model not only performs well on ASR, but also on accent identification when compared to separately-trained models. Given the recent success in end-to-end models [11, 122, 151–159], we use a bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) acoustic model trained with the connectionist temporal classification (CTC) loss function for acoustic modeling. The accent identification (AID) network is also a BLSTM, but includes an average pooling layer to compute an utterance-level accent embedding. We also introduce a joint architecture where the lower layers of the network are trained using AID as the auxiliary task while multi-accent acoustic modeling remains the primary task of the network.

Next, we use the AID network as a hard switch between the accent-specific output layers of the CTC AM. Preliminary experiments on the Wall Street Journal American English and Cambridge British English corpora demonstrate that our joint model with the AID-based hard-switch achieves lower WER when compared with the state-of-the-art multi-task AM. We also show that the AID model also benefits from joint training.

8.2 Related Work

The most closely related work to ours is from [144], which illustrated that hierarchical grapheme-based AM with auxiliary phoneme-based AMs in four English dialects trained with CTC significantly outperformed accent-specific AMs and grapheme-based AM, respectively, while achieving competitive WER with phoneme-based multi-accent AM. Similarly, Yi et al [147] also trained a multi-accent phoneme-based AM with CTC loss, but instead, adapted accent-specific output layer using its target accent.

Other relevant work compared the performance of training accent or dialect specific acoustic models and joint models. These approaches predicted context-dependent (CD) triphone states using DNNs, and used a weighted finite state transducer (WFST)-based decoder. For example, senones on accents of Chinese are predicted by assuming all accents within a language share a common CD state inventory [145, 146]. Elfeky et al [143] implemented a dialectal multi-task learning (DMTL) framework on three dialects of Arabic using the prediction of a unified set of CD states across all dialects prediction as the primary task and dialect identification as the secondary task. DMTL model deviated from ours in that it directly predicted CD states using convolutional-LSTM-DNNs (CLDNN), and was trained with either cross-entropy or state-level minimum Bayes risk, while ignoring the secondary dialect identification output at recognition time. This DMTL model was trained on all dialectal data and underperformed the dialect-specific model. Dialectal knowledge distilled (DKD) model was also designed in [143], which achieved results competitive to, but below, dialect-specific models.

The effectiveness of dialect-specific models motivated investigations into how to use ensemble methods on multiple dialect-specific acoustic models for recognition. Soto et al [141] explored approaches of selecting and combining the best decoded hypothesis from a pool of dialectal recognizers. This work is still different from

ours in that we make perform selection directly using predicted dialect. Huang et al [8] used a similar strategy to ours by identifying accent first followed by acoustic model selection, however, this work only considered GMMs as the classifier.

8.3 Method

Our proposed system consists of multiple accent-specific acoustic models and accent identification model. We will describe these components and their joint model in this section. Acoustic model selection based on the hard-switch between accent-specific models is illustrated in Section 8.3.4.

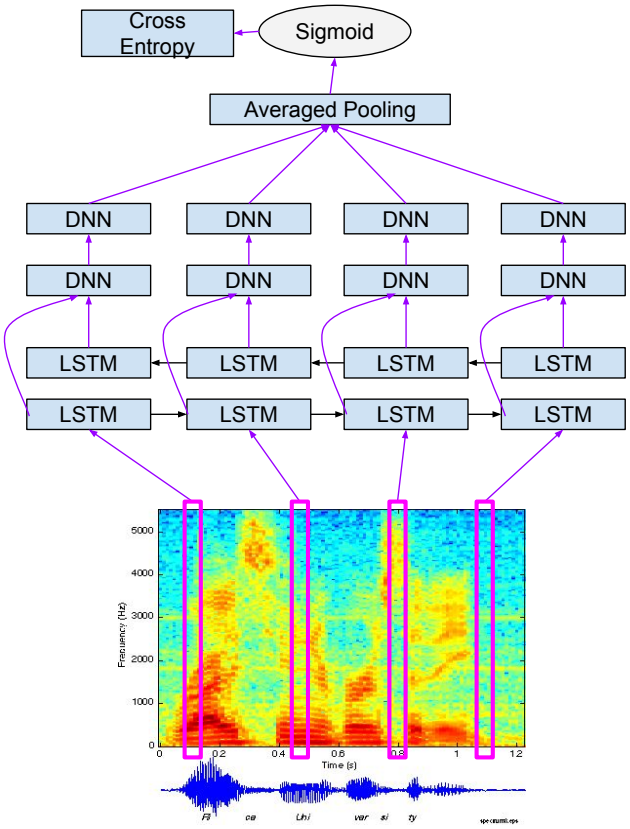


Figure 8.1: Proposed accent identification (AID) model with BLSTMs and average-pooling.

8.3.1 Accent Identification

Accurate identification of a speaker's accent is essential to the pipelined ASR systems, since accent identification (AID) errors can cause large mismatch to acoustic models. Given the hypothesis that accents can be discriminated by spectral features, researchers have attempted to model the spectral distribution of each accent using GMMs. Recently, DNNs have been explored as a much more expressive model compared to GMMs, especially in modeling probability distributions.

We implemented an independent AID that summarizes low-level acoustic features of an utterance by a stack of bidirectional LSTMs (BLSTMs) and DNN projection layers. An average-pooling layer is applied on top of transformed acoustic features, because the acoustic realization of a speaker's accent may not be observable in each frame. Applying average-pooling gives us a more robust estimate of accent-dependent acoustic features. We note that we assume that the speaker's accent is fixed over the entire utterance.

Figure 8.1 depicts details of this AID model. A single sigmoidal neuron is used at the output layer for classification because we are only classifying between accents of English—US and UK. We trained the AID network using the cross-entropy loss.

8.3.2 Multi-Accent Acoustic Modeling

Recently, end-to-end (E2E) systems have achieved comparable performance to traditional pipelined systems such as hybrid DNN-HMM systems. These E2E systems come with the benefit of avoiding time-consuming iterations between alignment and model building. RNNs using the CTC loss function are a popular approach to E2E systems [151]. The CTC loss computes the total likelihood of the output label sequence given the input acoustics over all possible alignments. It achieves this by introducing a special *blank* symbol that augments the label sequence to make its length equal to the length of the input sequence. Clearly, there are multiple such augmented sequences, and CTC uses the forward-backward algorithms to efficiently sum the likelihoods of such sequences. The CTC loss is

$$p(\mathbf{l}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{l})} p(\pi|\mathbf{x}) \tag{8.1}$$

where $\mathbf{l}$ is the output label sequence, $\mathbf{x}$ is the input acoustic sequence, π is a blank-augmented sequence for $\mathbf{l}$, and $\mathcal{B}^{-1}(\mathbf{l})$ is the set of all such sequences. During

decoding, the target label sequences can be obtained by either greedy search or a WFST-based decoder.

Our multi-accent acoustic model combines two CTC-based AMs, one for each accent. We applied multiple BLSTM layers shared by two accents to capture accent-independent acoustic features, and placed separate DNNs for each AM to extract accent-specific features. Figure 8.2 describes the structure of multi-accent acoustic model. Both AMs are jointly trained with an average of the two accent-specific CTC losses.

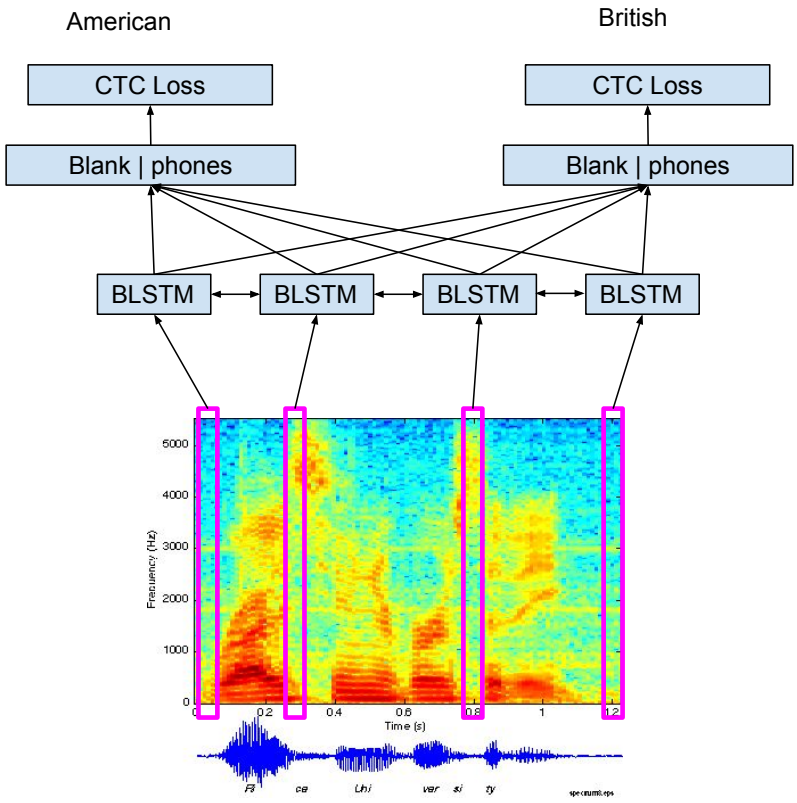


Figure 8.2: This figure shows the multi-accent acoustic model.

At test time, this multi-accent model requires knowledge of the speaker's accent to pick out of the two accent-specific targets. We experimented with both the oracle accent label, and using a trained AID network to make this decision.

8.3.3 Joint Acoustic Modeling with AID

The previous multi-accent model assumes that multi-tasking between the phone sets of the two accents is sufficient to make the network learn accent-specific acoustic

information. An alternate approach is to explicitly supervise the network with accent information. This leads us to our joint model, with multi-accent acoustic modeling as primary tasks at higher layers, and with AID as an auxiliary task at lower level layers, as shown in Figure 8.3. This joint model aggregates two modules with the same structures to the aforementioned models in Section 8.3.1 and Section 8.3.2, and can be jointly trained in an end-to-end fashion with the objective function,

$$\min_{\Theta} \mathcal{L}_{\text{Joint}}(\Theta) = (1-\alpha) * \mathcal{L}_{\text{AM}}(\Theta) + \alpha * \mathcal{L}_{\text{AID}}(\Theta)$$

where α is an interpolation weight balancing between CTC loss of multi-accent AMs and the cross-entropy loss of AID, and Θ is the model parameters. CTC loss $\mathcal{L}_{\text{AM}}$ sums up the probabilities of all possible paths corresponding to Equation (8.1), while AID classification loss $\mathcal{L}_{\text{AID}}$ is cross-entropy. The two losses are at different scales, so the optimal value of α needs to be tuned on development data.

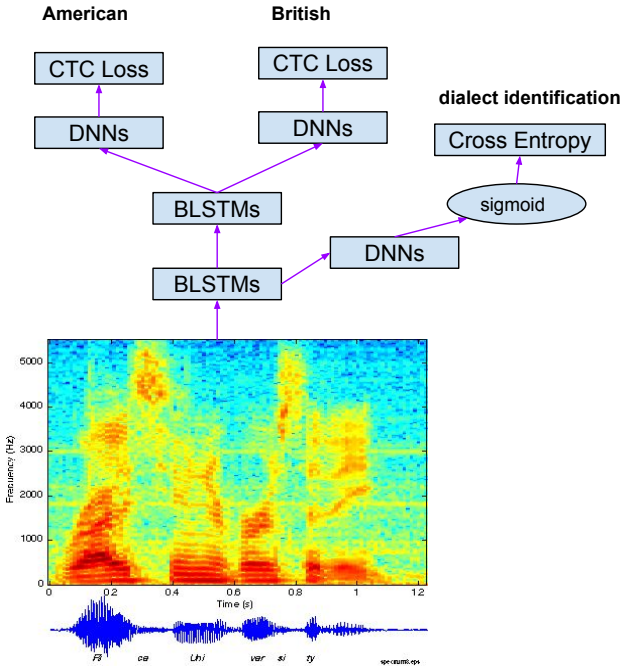


Figure 8.3: Proposed joint model for accent identification and acoustic modeling.

8.3.4 Model Selection by Hard-Switch

Given a trained CTC-based multi-accent acoustic model and AID classifier, we apply maximum likelihood estimation to switch between the accent-specific output layers $\mathbf{y}_{\text{US}}$ and $\mathbf{y}_{\text{UK}}$. Let $P_{\text{AID}}\left(\text{US}|\mathbf{x}\right)$ denote the probability of the US accent estimated by AID. We threshold this probability at 0.5 to obtain the accent hard-switch $s_{\text{AID}}\left(\text{US}|\mathbf{x}\right)$. Hence, we pick the output layer as Equation (8.2) shows. We note that this strategy applies to both the multi-accent model and the joint model.

$$\mathbf{y} = \begin{cases} \mathbf{y}_{\text{US}} & \text{if } s_{\text{AID}}(\text{US}|\mathbf{x}) = 1 \\ \mathbf{y}_{\text{UK}} & else \end{cases} \tag{8.2}$$

8.4 Experiments

We perform experiments on two dialects of English corpora—Wall Street Journal American English and Cambridge British English. They contain overlapping but distinct phone sets of 42 and 45 phones respectively. Both corpora contain approximately 15 hours each of audio. We held-out 5% of the training data as a development set. The window size of each speech frame is 25ms with a frame shift of 10ms. We extracted 40-dimensional log-Mel scale filterbanks and performed per-utterance cepstral mean subtraction. We did not use any vocal tract length normalization. We then stacked neighboring frames and picked every alternate frame to get a 80-dimensional acoustic feature stream at half the frame rate. Various models are compared in terms of phone error rate (PER) and word error rate (WER). Particularly, we obtain the PER after simple frame-wise greedy decoding from the DNN projection outputs after removing repeated phones and the *blank* symbol. The Attila toolkit [160] is used to report WER by applying WFST-based decoding. Evaluation is performed on `eval93`[1] American English and `si_dt5b`[2] British English.

Our joint model uses four BLSTM layers where the lowest layer is attached to the AID network and the highest single layer connects to two accent-specific softmax layers. A single DNN layer with 320 hidden units is used for each task. The weights for all models are initialized uniformly from $[-0.01, 0.01]$. Adam optimizer [138] with initial learning rate $5\text{e}{-4}$ is used, and the gradients are clipped

[1] `catalog.ldc.upenn.edu/ldc93s6a`
[2] `catalog.ldc.upenn.edu/LDC95S24`

to the range $[-10, 10]$. We discard the training utterances that are longer than 2000 frames. New-Bob annealing [161] on the held-out data is used for early stopping, where the learning rate is cut in half whenever the held-out loss does not decrease. For the purpose of fair comparison, we used a four layer BLSTM for the baseline acoustic models as well.

Various models are briefly described as follows:

- ASpec: phoneme-based accent-specific AMs that are trained separately on mono-accent data.
- MTLP: phoneme-based multi-accent AMs that are jointly trained on two accents.
- Joint: proposed phoneme-based joint acoustic model with AID.

8.4.1 Empirical weights for balancing different losses

Our joint model is sensitive to the interpolation weight α between the AM CTC loss and AID cross-entropy loss. We tuned α on development data. Figure 8.4 depicts relationship between overall PER of two accents and different α values. When α goes larger, overall PER increases but with small fluctuations, especially at α of 0.01 and 0.2. The PER tends to be the largest if α is 1.0, which is expected since the weights of neural networks are updated only using the AID errors. We found the optimal value of α to be 0.001, which achieved minimum PER of 12.02%. Figure 8.5 illustrates the trend of AID accuracy over different α values. Weights between 0.001 and 0.8 all perform well with accuracies greater than 92%, while tail values lead to even worse performance. When α is 0.5 and 0.005, the best performance is achieved with 97.77% accuracy.

8.4.2 Oracle performance for multi-accent acoustic models

We first evaluate the oracle performance of various models in Table 8.1. These results assume that the correct accent of each utterance is provided for all models. In other words, the acoustic model corresponds to the correct accent, i.e. the relevant target accent-specific softmax layer is used. It can be seen that the proposed joint model significantly outperforms the accent-specific model (ASpec) by 14.98% relative improvement in WER, and multi-task accent model (MTLP) by 8.81%. This observation indicates that deep BLSTM layers shared with multiple accent

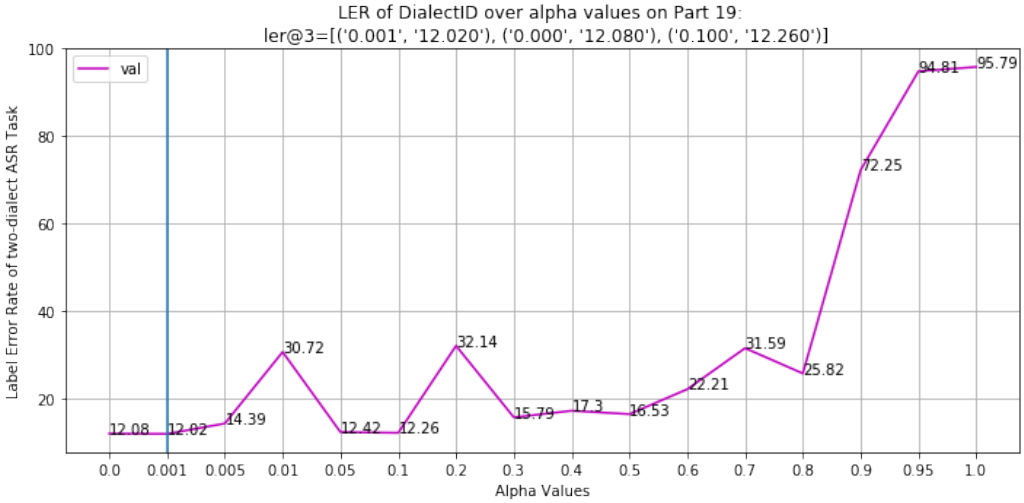


Figure 8.4: PER of joint acoustic model over AID loss weights α.

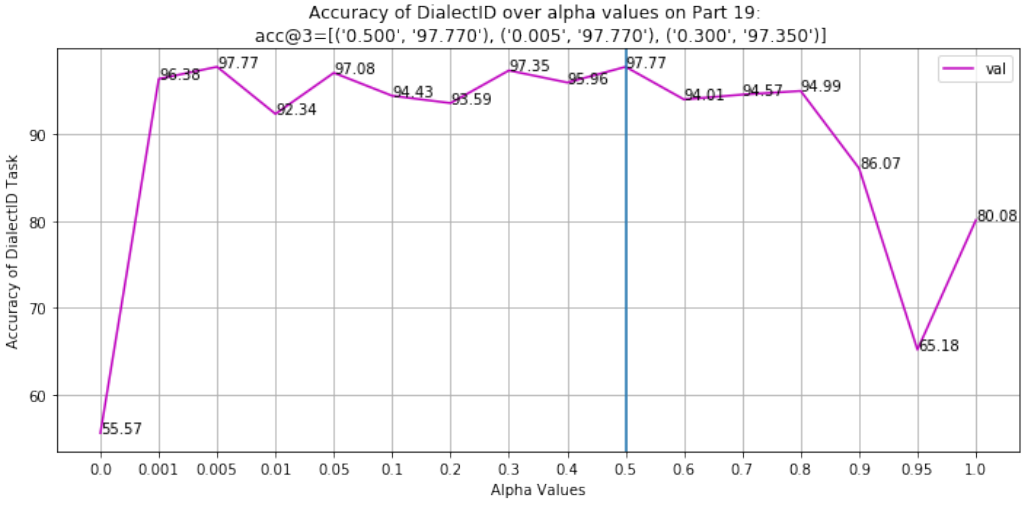


Figure 8.5: Accuracy of AID over various AID weights α.

AMs can learn expressive accent-independent features that refine accent-specific AMs. The auxiliary task, accent identification, also helps by introducing extra accent-specific information. The advantage of augmenting general acoustic features with specific information both implicitly learned by our joint model is observed in natural language processing [162] tasks as well. The value of implicit feature augmentation is a rich area for future investigation.

Table 8.1: Oracle performance that assumes that the true accent ID is known in advance. *PER* is phone error rate computed by greedy decoding; *WER* is word error rate after decoding with a WFST-graph incorporating a LM.

corpus	ASpec		MTLP		Proposed Model	
	PER	WER	PER	WER	PER	WER
British	17.68	11.20	17.29	10.20	14.48	9.50
American	15.95	9.50	14.75	9.10	11.93	8.10

8.4.3 Hard-switch using distorted AID

The oracle experiments in Section 8.4.2 demonstrate the value of our proposed joint model and the MTLP model when the AID classifier operates perfectly. This section demonstrates the impact of imperfect AID on the performance using hard-switch. Table 8.2 shows the results. Given a well-trained independent AID (ind. AID), our joint model still significantly outperforms the two baseline models, and MTLP achieves better WER than ASpec. In comparison to oracle WERs of all models, British WERs are relatively constant without any distortion, however, American English WERs deteriorate accordingly. This is because independent AID has 100% recall for British English utterances on the test data.

It is interesting to note that the biggest improvement over ASpec in WER comes when using the joint model (16.35%) instead of the MTLP model (6.73%) with an independent AID model. The improvement upon further using the AID from the joint model itself is still larger (17.31%). This indicates that the joint model has already learned sufficient accent-specific information through the accent supervision in the lower layers.

Table 8.2: WERs of hard-switch using distorted AID. The *rel.* shows the relative improvement over ASpec; *ind. AID* applies an independent neural AID trained separately. Our *Proposed Model* applies the AID jointly learn with multi-accent AMs.

Corpus	Pipelines with ind. AID			Proposed Model (rel.)
	ASpec	MTLP (rel.)	Joint (rel.)	
British	11.2	10.2 (-8.93)	9.5 (-15.18)	9.5 (-15.18)
American	10.4	9.7 (-6.73)	8.7 (-16.35)	8.6 (-17.31)

8.5 Conclusion

This chapter studies state-of-the-art approaches of acoustic modeling across multiple accents. We note that these prior approaches do not explicitly include accent information during training, but do so only indirectly, for example through the different phone inventories for various accents. We propose an end-to-end multi-accent acoustic modeling approach that can be jointly trained with accent identification. We use BLSTM-RNNs to design acoustic models that can be trained with CTC, and apply an average pooling to compute utterance-level accent embedding. Experiments show that both multi-accent acoustic models and accent identification benefit each other, and our joint model using hard-switch outperforms the state-of-the-art multi-accent acoustic model baseline with a separately-trained AID network. We obtain a 6.86% relative improvement in WER on British English, and 11.34% on American English.

CHAPTER 9

WHEN CTC TRAINING MEETS ACOUSTIC LANDMARKS

Connectionist temporal classification (CTC) provides an end-to-end acoustic model training strategy. CTC learns accurate AMs without time-aligned phonetic transcription, but sometimes fails to converge, especially in resource-constrained scenarios. In this chapter, the convergence properties of CTC are improved by incorporating acoustic landmarks. We tailored a new set of acoustic landmarks to help CTC training converge more rapidly and smoothly while also reducing recognition error. We leveraged new target label sequences mixed with both phone and manner changes to guide CTC training. Experiments on TIMIT demonstrated that CTC based acoustic models converge significantly faster and smoother when they are augmented by acoustic landmarks. The models pretrained with mixed target labels can be further finetuned, resulting in a phone error rate 8.72% below baseline on TIMIT. Consistent performance gain is also observed on WSJ (a larger corpus) and reduced TIMIT (smaller). With WSJ, we are the first to succeed in verifying the effectiveness of acoustic landmark theory on a mid-sized ASR task.

9.1 Introduction

Automatic speech recognition (ASR) is a sequence labeling problem that translates a speech waveform into a sequence of words. Recent success of hidden Markov model (HMM) combined with deep neural networks (DNNs) or recurrent neural networks has achieved a word error rate (WER) on par with human transcribers [3, 163]. These hybrid acoustic models (AMs) are typically optimized by cross-entropy (CE) training which relies on accurate frame-wise context-dependent state alignments pre-generated from a seed AM. The connectionist temporal classification (CTC) loss function [116], in contrast, provides an alternative method of AM training in an end-to-end fashion—it directly addresses the sequence labeling problem without prior frame-wise alignments. CTC is capable of learning to con-

struct frame-wise paths implicitly bridging between the input speech waveform and its context-independent target, and it has been demonstrated to outperform hybrid HMM systems when the amount of training data is large [11, 122, 152]. However, its performance degrades and is even worse than traditional CE training when applied to small-scale data [164].

Training CTC models can be time-consuming and sometimes models are apt to converge to even a sub-optimal alignment, especially on resource-constrained data. In order to alleviate such common problems of CTC training, additional tricks are needed, for example, ordering training utterances by their lengths [11] or bootstrapping CTC models with models CE-trained on fixed alignments [165]. The success of bootstrapping with prior alignments indicates that external phonetic knowledge may help to regularize CTC training towards stable and fast convergence. Furthermore, another investigation [16] reveals that the spiky predictions of CTC models tend to overlap with the vicinity of acoustic landmarks where abrupt manner changes of articulation occur [31]. The possible coincidence of CTC peaks overlapping acoustic landmarks suggests a number of possible approaches for reducing the data requirements of CTC, including cross-language transfer (using the relative language-independence of acoustic landmarks [15]) and informative priors.

Many efforts have been attempted to augment acoustic modeling with acoustic landmarks [12, 13, 15] which are detected by accurate time-aligned phonetic transcriptions. To the best of our knowledge, only TIMIT [56] (5.4 hours) provides such fine-grained transcriptions. The value of testing these approaches are limited since the only available corpus is very small. It is worth further exploring the power of landmark theory when scaled up to large corpus speech recognition.

In this chapter, we propose to augment phone sequences with acoustic landmarks for CTC acoustic modeling and leverage a two-phase training procedure with pretraining and finetuning to address CTC convergence problems. Experiments on TIMIT demonstrate that our approaches not only help CTC models converge more rapidly and smoothly, but also achieve a lower phone error rate, up to 8.72% phone error rate reduction over CTC baseline with phone labels only. We also investigate the sensitivity of our approaches to the size of training data on subsets of TIMIT (smaller corpora) and WSJ [166] (a larger corpus). Our findings demonstrate that label augmentation generalizes to larger and smaller training datasets, and we believe this is the first work that applies acoustic landmark theory to a mid-sized ASR corpus.

9.2 Background

9.2.1 Connectionist Temporal Classification

Recent end-to-end systems have attracted much attention, for example, because they avoid time-consuming iterations between alignment and model building [116, 151]. The CTC loss computes the total likelihood of the target label sequence over all possible alignments given an input feature sequence, so that the computation is more expensive than frame-wise cross-entropy training. A blank symbol is introduced to compensate for the difference in length between an input feature sequence and its target label sequence. Forward-backward algorithms are used to efficiently sum the likelihood over all possible alignments. The CTC loss is defined as,

$$\mathcal{L}_{ctc} = -\log p\left(\mathbf{y}|\mathbf{x}\right) = -\log \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{y})} p(\boldsymbol{\pi}|\mathbf{x})$$

where $\mathbf{x}$ is an input feature sequence, $\mathbf{y}$ is the target label sequence of $\mathbf{x}$, $\boldsymbol{\pi}$ is one of blank-augmented alignments of $\mathbf{y}$, and $\mathcal{B}^{-1}(\mathbf{y})$ calculates the set of all such alignments. During decoding, the n-best list of predicted label sequences can be achieved by either a greedy search or a beam search based on weighted finite state transducers (WFSTs). In the following experiments, our acoustic models are trained by the phoneme CTC loss, and we report phone error rates on TIMIT (a smaller corpus) through an one-best greedy search and word error rates on WSJ (a larger corpus) through an one-best WFSTs beam search, respectively.

9.2.2 Acoustic Landmarks

Acoustic landmark theory originates from experimental studies of human speech production and speech perception. It claims there exist instantaneous acoustic events that are perceptually salient and sufficient to distinguish phonemes [31]. Automatic landmark detectors can be knowledge-based [43] or learned [44]. Landmark-based ASR has been shown to slightly reduce the WER of a large-vocabulary speech recognizer, but only in a rescoring paradigm using a very small test set [44]. Landmarks can reduce computational load for DNN/HMM hybrid models [12, 13] and can improve recognition accuracy [15]. Previous works [12, 13, 15, 17] annotated landmark positions mostly following experimental findings presented in [25, 40]. Four different landmarks are defined to capture

positions of vowel peak, glide valley in glide-like consonants, oral closure and oral release.

9.3 Methods

9.3.1 Distinctive Features and Landmark Definition

Distinctive features (DFs) concisely describe sounds of a language at a sub-segmental level, and they have direct relations to acoustics and articulation. These features take on binary encodings of perceptual, phonological, and articulatory speech sounds [167]. A collection of these binary features can distinguish each segment from all others in a language. Autosegmental phonology [168] also suggests that DFs have an internal organization with a hierarchical relationship with each other. We follow these linguistic rules to select two primary features—*sonorant* and *continuant*—that distinguish among the manner classes of articulation, resulting in a four-way categorization shown in Table 9.1. We define landmarks as the changes in the value of one of these two distinctive features where the TIMIT phone inventory is applied. The standard phoneme set used by WSJ ignores detailed annotations of oral closures, for example /bcl/, so that we merge together [-,+*continuant*] features under [-*sonorant*] column in Table 9.1, resulting in a three-way categorization for WSJ experiments instead.

Table 9.1: Broad classes of sounds on TIMIT.

Manner	-sonorant	+sonorant
-continuant	bcl dcl gcl kcl pcl q tcl	em en eng m n ng
+continuant	b d g k p t ch jh dh f hh hv s sh th v z zh	aa ae ah ao aw ax ax-h axr ay dx eh el ey ih ix iy l nv ow oy r uh uw ux w y er

9.3.2 Augmenting Phone Sequences With Landmarks

We defined two methods of augmenting phone label sequences with acoustic landmarks. *Mixed Label 1* only inserts landmarks between two broad classes of sounds where manner changes occur; *Mixed Label 2* inserts landmarks between phones even if manner changes don't exist. Figure 9.1 demonstrates an example of our two augmentation methods.

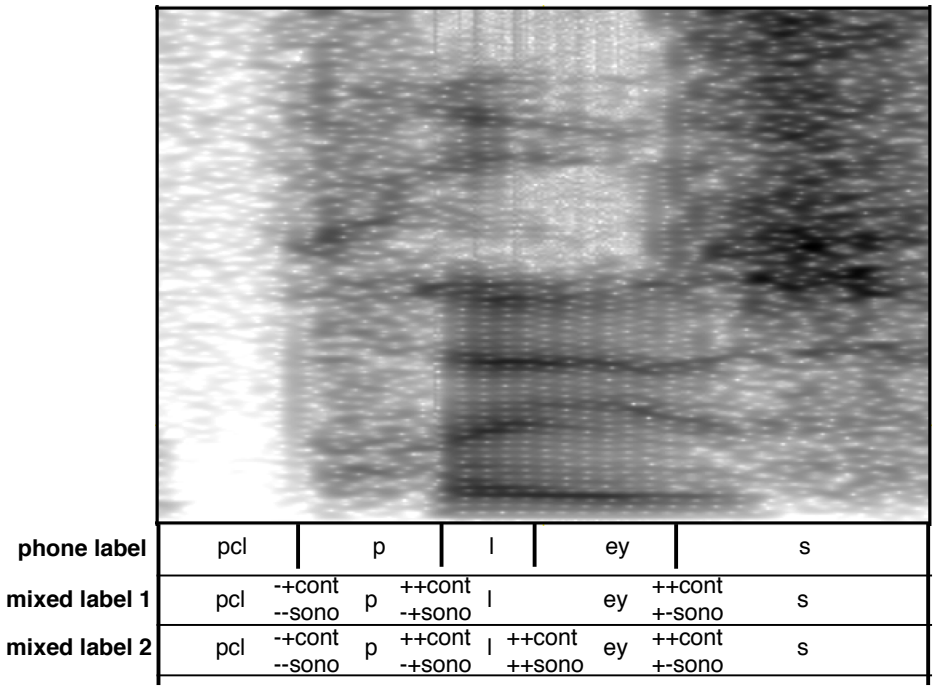


Figure 9.1: Examples of target label sequences for the word "PLACE". The audio clip is selected from `SI792` on TIMIT.

CTC only requires a single target label sequence, so that augmenting phone sequences with landmarks can relax the need for time-aligned phone transcriptions. With a blank label present between two phones in the training target sequence, the vanilla CTC training can be considered as already experimenting with the scenario where a dedicated phone boundary label is added to the label set. CTC is thus an ideal baseline for our experiments.

9.3.3 Acoustic Modeling using CTC

We follow a pretraining and finetuning procedure to train our CTC models. At the phase of pretraining, the AM initializes weights randomly and is trained by one of

our mixed label sequences until convergence; at the phase of finetuning, the AM initializes weights from the pretrained model and continues to be trained by a label sequence with only phones. These two phases of training take the same acoustic features. Figure 9.2 briefly illustrates the whole procedure. The top output layer calculates a posterior distribution over symbols combined with both phones and landmarks, while the bottom output layer calculates it over only phones.

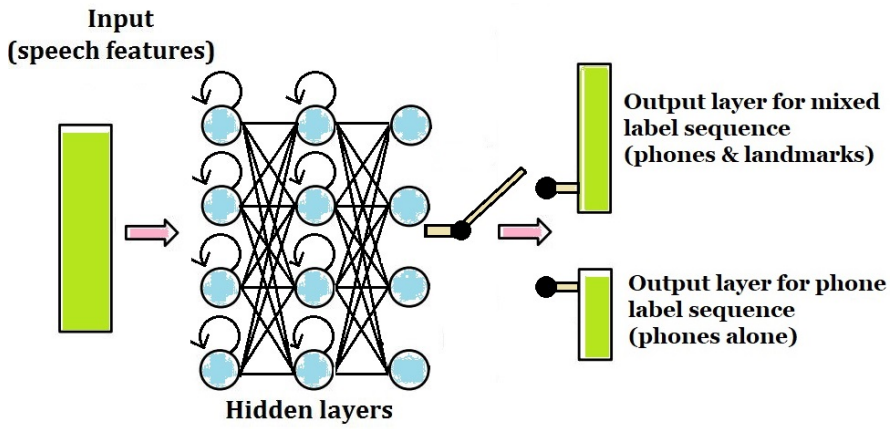


Figure 9.2: Two-phase acoustic modeling: top output layer pretrains with mixed labels and bottom output layer finetunes with phone labels only.

9.4 Experiments

9.4.1 Configurations

We conducted our experiments on both the TIMIT [56] and WSJ [166] corpora. We used 40-dimensional log mel filterbank energy features computed with 10ms shift and 20ms span. No delta features or frame stacking were used. The recurrent neural networks stacked two layers of bidirectional LSTMs, each with 1024 cells (512 cells per direction), capped by a fully connected layer with 256 neurons. Weights are initialized randomly from Xavier uniform distribution [169]. New-Bob annealing [170] is used for early stopping after a minimum waiting period of two epochs. The initial learning rate is 0.0005. The TIMIT baseline is trained on 61 phones. The WSJ baseline is trained on 39 phones[1] defined in the CMU pronunciation dictionary. One-best greedy search is applied to calculate the phone

[1] `https://github.com/Alexir/CMUdict/blob/master/cmudict-0.7b.phones`

error rate (PER). We did not map TIMIT phones to CMU phone set (39 phones). In order to make a fair comparison, all baselines went through the same two-phase training with pretraining and finetuning. One-best beam search based on WFSTs is applied to calculate the word error rate in WSJ experiments using decoding graphs with a primitive trigram (tg) and pruned trigram (tgpr) from EESEN[2]. We use the same train/dev/test split from Kaldi Recipes for TIMIT and WSJ.

9.4.2 Experiments on TIMIT

Figure 9.3 presents the development set PER as a function of training epoch. The PER for mixed sequence represented by the red and yellow lines in Figure 9.3 is calculated after landmark labels have been removed from the output sequence. In the pretrain phase, models trained on augmented labels do not seem to have any advantage in terms of error rate. However, the models converge much more rapidly and smoothly. After pretraining, both the baseline and mixed-label systems are finetuned; the mixed-label system (purple line in Fig. 9.3) returns a model that is more accurate.

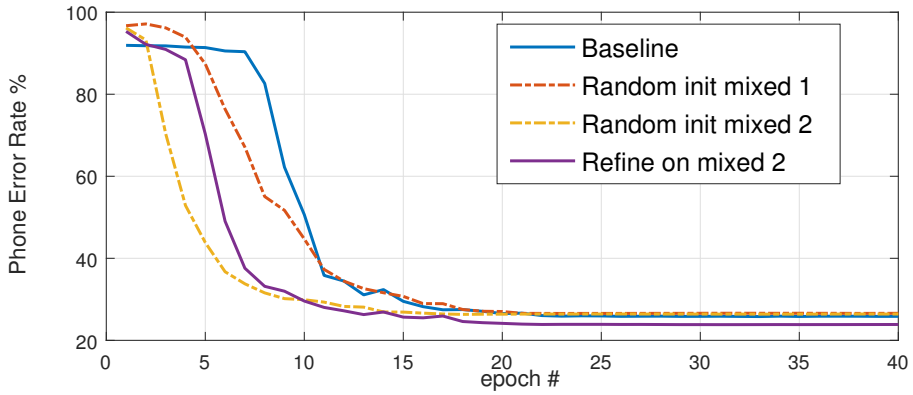


Figure 9.3: PER as a function of training epoch. PER is calculated against only phones after landmarks are removed.

The exact PERs for different setups on the TIMIT test set are reported in Table 9.2. Our baseline achieved a PER of 30.36%, which was not improved by finetuning. This is higher than PER reported elsewhere (e.g., [116]), because nobody else calculates PER on the full TIMIT set of 61 phones. As shown in Table 9.2, if we train with mixed labels and strip away landmarks from the

[2] https://github.com/srvk/eesen/blob/master/asr_egs/wsj/run_ctc_phn.sh

hypothesis sequence, landmarks provide little benefit. However, the *Mixed 1* and *Mixed 2* systems achieved lower PERs after the finetuning stage by 4.64% and 8.72% relative reduction, respectively. Apparently, a phone sequence augmented with landmarks can be learned more accurately than a raw phone sequence, perhaps because the acoustic features of manner transitions are easy to learn, and help to time-align the training corpus. The *Mixed Label 2* set outperforms *Mixed Label 1*, apparently because the extra boundary information in *Mixed Label 2* is beneficial to the training algorithm.

Table 9.2: Comparison between baseline and our proposed models with augmented target labels in PER (%). The number in the parentheses denotes the relative reduction over baseline.

	Baseline	Mixed 1	Mixed 2
random init	30.36	30.98	29.10
finetuned	30.36	**28.96** (4.64%)	**27.72** (8.72%)

It is not clear why a finetuning stage is needed in order for *Mixed 1* to beat the baseline. One possibility is that landmark labels are helpful for some tokens, and harmful for others; pretraining uses the helpful landmarks to learn better phone alignments, then finetuning permits the network to learn to ignore the harmful landmark tokens. We looked into the prior distribution on TIMIT, presented in Figure 9.4, of both phones (top subplot, with phones ordered in the same way as they occurred in Table 9.1) and landmarks (bottom subplot, *Mixed Label 2* ordered in category permutation using *continuant* as the first variable and *sonorant* as the second). The table reveals that the distribution of landmarks is not balanced. Most labels indicate a transition related to the [+*continuant,*+*sonorant*] phones. A skewed landmark support is not ideal for augmenting phone recognizer training as it tends to provide the same and redundant information for many training sequences.

9.4.3 Datasets Smaller and Larger than TIMIT

To solidify our findings, we further investigated the sensitivity of our approaches to the size of training data on subsets of TIMIT (smaller corpora) and WSJ (a larger corpus). In this section, we only demonstrate the experiments using *Mixed Label 2* augmentation method since it outperforms *Mixed Label 1* in the previous

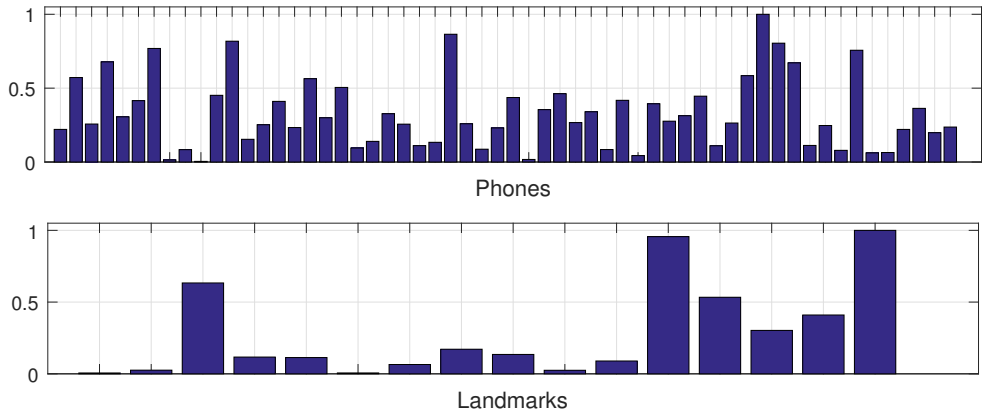


Figure 9.4: Prior distributions of phones and acoustic landmarks.

discussion. We report PER/WER results for finetuned models.

Figure 9.5 shows the PER results by stretching the amount of training data on TIMIT. Both the proposed model and baseline fail to converge when 75% of the training data is used. We observe that both models start to predict a constant sequence (usually made up of two to three most frequent phones) for all utterances. Scheduled reducing the learning rate by New-Bob annealing can't help to converge to an optimal. Increasing the amount of training data helps both models converge. The baseline needs 90% of TIMIT to converge, while the proposed system only needs 80% of TIMIT.

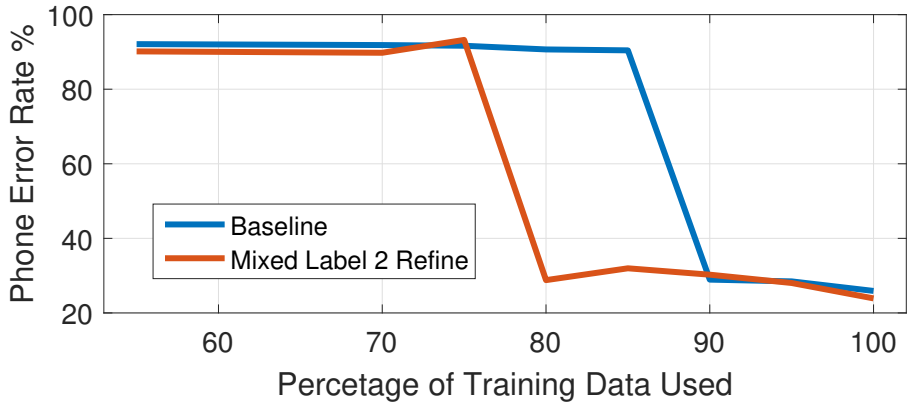


Figure 9.5: PERs by stretching the amount of training data on TIMIT.

When scaling up to a even larger corpus on WSJ, the proposed *Mixed Label 2* system could achieve better performance over the baseline consistently in terms of all metrics as shown in Table 9.3. Our baseline system slightly under-performs the results published in EESEN [122] because our network is shallower and the

acoustic inputs do not include any dynamic (delta) features, but the benefit of the proposed landmark augmentation method still applies. To our knowledge, this is the first work to show that manner-change acoustic landmarks reduce both PER and WER on a mid-sized ASR corpus.

Table 9.3: Label Error Rate (%) on WSJ, where tg and tgpr denote decoding graphs with primitive and pruned trigrams.

	PER		WER (tgpr/tg)	
	eval92	dev93	eval92	dev93
Baseline	8.70	12.38	8.75/8.17	13.15/12.31
Mixed 2	**8.12**	**11.49**	**8.35**/8.19	**12.86/12.28**

9.5 Conclusion

We proposed to augment CTC with acoustic landmarks. We modified the classic landmark definition to suit the CTC criterion and implemented a pretraining-finetuning training procedure to improve CTC AMs. Experiments on TIMIT and WSJ demonstrated that CTC training becomes more stable and rapid when phone label sequences are augmented by landmarks, and achieves a significantly lower (8.72% relative reduction) asymptotic PER. The advantage is consistent across corpora (TIMIT, WSJ) and across metrics (PER, WER). CTC with landmarks converges when the dataset is too small to train the baseline, and it also converges without the need of time alignments on a mid-sized standard ASR training corpus (WSJ).

CHAPTER 10

CONCLUSION

10.1 Summary

This dissertation investigated both knowledge-gnostic and knowledge-agnostic approaches about acoustic and articulatory phonetics in order to address the issues of linguistic mismatches for current automatic speech recognition systems.

- **Knowledge-gnostic approaches:** acoustic landmarks exploit quantal non-linear articulator-acoustic relationships which identify times when acoustic patterns of linguistically motivated distinctive features are mostly salient. Acoustic cues extracted in the vicinity of landmarks are demonstrated to contain more information for identification of articulator manner changes and for classification of distinctive features than the cues extracted from other times in the signal. Further ASR decoding experiments were conducted on TIMIT by a heuristic method of weighting acoustic likelihood scores for speech frames, and we observed that landmark speech frames are more informative for recognition than other frames as expected. In consideration of the superb property of informative landmark frames, we also validated the power of portability of English landmark theory to other languages, for example, Chinese, in the application of pronunciation error detection tasks. In order to deal with scarce landmark transcripts in any languages other than English, two landmark detectors—deep neural model and spiky CTC model—are implemented so that accurate landmark transcripts are accessible to any languages. Acoustic landmark knowledge also benefits end-to-end acoustic modeling using CTC, such that CTC training could converge more rapidly and smoothly and achieve a lower word error rate. These contributions provide a solid foundation for many applications: segmenting phonemes in low-resourced languages for fine-grained analysis; predicting more accurate manner and place of articulators that can bridge

mismatches across languages; altogether moving further toward the holy grail of language-independent acoustic phonetic speech recognizer.

- **Knowledge-agnostic approaches:** human listeners can well perceive voices with linguistic mismatches although they may not have any backgrounds relevant to acoustic and articulatory phonetics. Human typically memorize detailed mental representations of phonological forms and traces of individual voices or type of voices implicitly. Imitation learning directly from the ability of human perception rely on a large scale recordings while disregards any background knowledge. We proposed to link the training of acoustic and accent altogether in a manner similar to the learning process in human speech perception. We showed that this joint model not only performed well on ASR with multiple accents, but also on accent identification when compared to separately-trained models.

10.2 Future Directions

Recent success on automatic speech recognition attracts many experts with diverse backgrounds including speech science, machine learning, and deep learning, and they help to push the progress of ASR techniques to a new cutting edge. Researchers are so obsessed with statistical modeling methods (e.g. word-pieces end-to-end CTC framework and attention-based encoder-decoder modeling) on large scale speech recordings that less efforts have been made to leverage the benefits of acoustic and articulatory phonetics knowledge. One of my work on knowledge-agnostic joint modeling of acoustics and accents falls in this category as well, however, it is still a long row to hoe for commercial products deployment in real-world tasks since the training data can't guarantee to enumerate all possible varieties of dialects or languages. Collecting enough data is an ideal solution but it is way too expensive and time-consuming. This dissertation re-visited classic theories relevant to linguistic prior knowledge, and experiments on speech tasks demonstrated the potentials that distinctive feature classifiers anchored in the vicinity of landmarks could help to build a bridge for dealing with mismatches across different local or global varieties in a dialect continuum. The discoveries in these contributions indicate another promising research direction, that is to combine both acoustics and articulatory phonetics knowledge into the state-of-the-art statistical

ASR models. It is the time for the oft-quoted statement to change: “Every time I ~~fire~~ *hire* a linguist, the performance of the speech recognizer goes up”.

REFERENCES

[1] Abdel-Rahman Mohamed, George Dahl, and Geoffrey Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009, vol. 1, p. 39.

[2] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[3] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, and Phil Hall, "English conversational telephone speech recognition by humans and machines," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017, pp. 132–136.

[4] Vladimir Vapnik, *The nature of statistical learning theory*, Springer Science & Business Media, 2013.

[5] Leslie G Valiant, "A theory of the learnable," in *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*. ACM, 1984, pp. 436–445.

[6] Jan D. ten Thije and Ludger Zeevaert, *Receptive multilingualism: linguistic analyses, language policies and didactic concepts*, vol. 6, John Benjamins Publishing, 2007.

[7] Jack K Chambers and Peter Trudgill, *Dialectology*, Cambridge University Press, 1998.

[8] Chao Huang, Tao Chen, and Eric Chang, "Accent issues in large vocabulary continuous speech recognition," *International Journal of Speech Technology*, vol. 7, no. 2, pp. 141–153, 2004.

[9] Lori F Lamel and Jean-Luc Gauvain, "Cross-lingual experiments with phone recognition," in *Proceedings of the 18th IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1993, vol. II, pp. 507–510.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[11] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. ACM, 2016, pp. 173–182.

[12] Di He, Boon Pang Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Acoustic landmarks contain more information about the phone string than other frames for automatic speech recognition with deep neural network acoustic model," *The Journal of the Acoustical Society of America (JASA)*, vol. 143, no. 6, pp. 3207–3219, 2018.

[13] Di He, Boon Pang P Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Selecting frames for automatic speech recognition based on acoustic landmarks," *The Journal of the Acoustical Society of America (JASA)*, vol. 141, no. 5, pp. 3468–3468, 2017.

[14] Xuesong Yang, Xiang Kong, Mark Hasegawa-Johnson, and Yanlu Xie, "Landmark-based pronunciation error identification on Chinese learning," in *Proceedings of the 8th Biennial Meeting of the Speech Prosody Special Interest Group of the International Speech Communication Association (Speech Prosody)*. ISCA, 2016, pp. 247–251.

[15] Di He, Boon Pang Lim, Xuesong Yang, Mark Hasegawa-Johnson, and Deming Chen, "Improved ASR for under-resourced language through multi-task learning with acoustic landmarks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2018.

[16] Chuanying Niu, Jinsong Zhang, Xuesong Yang, and Yanlu Xie, "A study on landmark detection based on ctc and its application to pronunciation error detection," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, Dec 2017, pp. 636–640.

[17] Xiang Kong, Xuesong Yang, Mark Hasegawa-Johnson, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel, "Landmark-based consonant voicing detection on multilingual corpora," *The Journal of the Acoustical Society of America (JASA)*, vol. 141, no. 5, pp. 3468–3468, 2017.

[18] Xuesong Yang, Kartik Audhkhasi, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, and Mark Hasegawa-Johnson, "Joint model of accents and acoustics for multi-accent speech recognition," in *Proceeding of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5989–5993.

[19] Di He, Xuesong Yang, Boon Pang Lim, Yi Liang, Mark Hasegawa-Johnson, and Deming Chen, "When CTC training meets acoustic landmarks," in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, (To Appear).

[20] Xuesong Yang, Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng, "End-to-end joint learning of natural language understanding and dialogue manager," in *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5690–5694.

[21] Xuesong Yang, Anastassia Loukina, and Keelan Evanini, "Machine learning approaches to improving pronunciation error detection on an imbalanced corpus," in *Proceedings of the 5th IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 300–305.

[22] Kenneth N Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds., pp. 51–56. McGraw-Hill, New York, 1972.

[23] Kenneth N Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, Victoria A. Fromkin, Ed., pp. 243–255. Academic Press, Orlando, Florida, 1985.

[24] Kenneth N Stevens, "On the quantal nature of speech," *Journal of phonetics*, vol. 17, no. 1, pp. 3–45, 1989.

[25] Kenneth N Stevens, Sharon Y Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu, "Implementation of a model for lexical access based on features," in *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP)*. ISCA, 1992, pp. 499–502.

[26] Sadaoki Furui, "On the role of spectral transition for speech perception," *The Journal of the Acoustical Society of America (JASA)*, vol. 80, no. 4, pp. 1016–1025, 1986.

[27] Ralph N Ohde, "The developmental role of acoustic boundaries in speech perception," *The Journal of the Acoustical Society of America (JASA)*, vol. 96, no. 5, pp. 3307–3307, 1994.

[28] Roman Jakobson, C Gunnar Fant, and Morris Halle, *Preliminaries to speech analysis: The distinctive features and their correlates*, MIT press, 1951.

[29] Noam Chomsky and Morris Halle, *The sound pattern of English*, Studies in language. Harper & Row, 1968.

[30] Kenneth N Stevens, Samuel Jay Keyser, and Haruko Kawasaki, "Toward a phonetic and phonological theory of redundant features," *Invariance and variability in speech processes*, pp. 426–449, 1986.

[31] Kenneth N Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America (JASA)*, vol. 111, no. 4, pp. 1872–1891, 2002.

[32] Katrin Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 1998, pp. 891–894.

[33] Katrin Kirchhoff, *Robust speech recognition using articulatory information*, Ph.D. thesis, University of Bielefeld, 1999.

[34] Katrin Kirchhoff, Gernot A Fink, and Gerhard Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[35] Karen Livescu, Ozgur Cetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Haggerty, Bronwyn Woods, Joe Frankel, Mathew Magimai-Doss, and Kate Saenko, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," in *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2007, vol. IV, pp. 621–624.

[36] Sebastian Stuker, Florian Metze, Tanja Schultz, and Alex Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*. ISCA, 2003, pp. 1033–1036.

[37] Florian Metze, *Articulatory features for conversational speech recognition*, Ph.D. thesis, Karlsruhe Institute of Technology, 2005.

[38] Arild Brandrud Næss, Karen Livescu, and Rohit Prabhavalkar, "Articulatory feature classification using nearest neighbors," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2011, pp. 2301–2304.

[39] Chin-Hui Lee, Mark A Clements, Sorin Dusan, Eric Fosler-Lussier, Keith Johnson, Biing-Hwang Juang, and Lawrence R Rabiner, "An overview on automatic speech attribute transcription (ASAT)," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2007, pp. 1825–1828.

[40] Mark Hasegawa-Johnson, "Time-frequency distribution of partial phonetic information measured using mutual information," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2000, pp. 133–136.

[41] Steven M Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.

[42] Shizhen Wang, Steven M Lulich, and Abeer Alwan, "Automatic detection of the second subglottal resonance and its application to speaker normalization," *The Journal of the Acoustical Society of America (JASA)*, vol. 126, no. 6, pp. 3268–3277, 2009.

[43] Sharlene A Liu, "Landmark detection for distinctive feature-based speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 100, no. 5, pp. 3417–3430, 1996.

[44] Mark Hasegawa-Johnson, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchhoff, Karen Livescu, Srividya Mohan, Jennifer Muller, Kemal Sonmez, and Tianyu Wang, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2005, vol. I, pp. 213–216.

[45] Andrew Wilson Howitt, "Vowel landmark detection," in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)*. ISCA, 1999, pp. 2777–2780.

[46] Jeung-Yoon Choi, *Detection of consonant voicing: A module for a hierarchical speech recognition system*, Ph.D. thesis, Massachusetts Institute of Technology, 1999.

[47] Jung-In Lee and Jeung-Yoon Choi, "Detection of obstruent consonant landmark for knowledge based speech recognition system," *The Journal of the Acoustical Society of America (JASA)*, vol. 123, no. 5, pp. 3330–3330, 2008.

[48] Sukmyung Lee and Jeung-Yoon Choi, "Vowel place detection for a knowledge-based speech recognition system," *The Journal of the Acoustical Society of America (JASA)*, vol. 123, no. 5, pp. 3330–3330, 2008.

[49] Jung-Won Lee, Jeung-Yoon Choi, and Hong-Goo Kang, "Classification of fricatives using feature extrapolation of acoustic-phonetic features in telephone speech," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2011, pp. 1261–1264.

[50] Jung-Won Lee, Jeung-Yoon Choi, and Hong-Goo Kang, "Classification of stop place in consonant-vowel contexts using feature extrapolation of acoustic-phonetic features in telephone speech," *The Journal of the Acoustical Society of America (JASA)*, vol. 131, no. 2, pp. 1536–1546, 2012.

[51] Partha Niyogi, Chris Burges, and Padma Ramesh, "Distinctive feature detection using support vector machines," in *Proceedings of the 24th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1999, vol. 1, pp. 425–428.

[52] Rahul Chitturi and Mark Hasegawa-Johnson, "Novel time domain multiclass SVMs for landmark detection," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2006, pp. 2354–2357.

[53] Sarah E Borys, "An SVM front-end landmark speech recognition system," M.S. thesis, University of Illinois at Urbana-Champaign, 2008.

[54] Kaizhi Qian, Yang Zhang, and Mark Hasegawa-Johnson, "Application of local binary patterns for SVM-based stop consonant detection," in *Proceedings of the 8th Biennial Meeting of the Speech Prosody Special Interest Group of the International Speech Communication Association (Speech Prosody)*. ISCA, 2016, pp. 1114–1118.

[55] Zhimin Xie and Partha Niyogi, "Robust acoustic-based syllable detection," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2006, pp. 1571–1574.

[56] John S Garofalo, Lori F Lamel, William M Fisher, Johnathan G Fiscus, David S Pallett, and Nancy L Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus cdrom," *Linguistic Data Consortium*, 1993.

[57] Aren Jansen and Partha Niyogi, "A hierarchical point process model for speech recognition," in *Proceedings of the 33rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 4093–4096.

[58] Amit Juneja, *Speech recognition based on phonetic features and acoustic landmarks*, Ph.D. thesis, University of Maryland at College Park, 2004.

[59] Harvey Fletcher and Wilden A Munson, "Loudness, its definition, measurement and calculation," *Bell Labs Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[60] Kenneth N Stevens, *Acoustic phonetics*, vol. 30, MIT press, 2000.

[61] Juha Iso-Sipila, "Speech recognition complexity reduction using decimation of cepstral time trajectories," in *Proceedings of the 10th European Signal Processing Conference*. IEEE, 2000, pp. 1–4.

[62] Vincent Vanhoucke, Matthieu Devin, and Georg Heigold, "Multiframe deep neural networks for acoustic modeling," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7582–7585.

[63] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, and Carolina Parada, "Personalized speech recognition on mobile devices," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5955–5959.

[64] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2013, pp. 2345–2349.

[65] Thomas F Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, 2008.

[66] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[67] Hynek Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," in *Proceedings of the 5th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2003, pp. 255–260.

[68] Sven EG Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America (JASA)*, vol. 39, no. 1, pp. 151–168, 1966.

[69] Laurence Gillick and Stephen J Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceeding of the 14th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1989, pp. 532–535.

[70] NAC Cressie and HJ Whitford, "How to use the two sample t-test," *Biometrical Journal*, vol. 28, no. 2, pp. 131–148, 1986.

[71] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2014, pp. 338–342.

[72] Bertrand Delgutte and Nelson YS Kiang, "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics," *The Journal of the Acoustical Society of America (JASA)*, vol. 75, no. 3, pp. 897–907, 1984.

[73] Preethi Jyothi and Mark Hasegawa-Johnson, "Acquiring speech transcriptions using mismatched crowdsourcing," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, 2015, pp. 1263–1269.

[74] Amit Juneja and Carol Espy-Wilson, "A novel probabilistic framework for event-based speech recognition," *The Journal of the Acoustical Society of America (JASA)*, vol. 114, no. 4, pp. 2335–2335, 2003.

[75] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[76] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4460–4464.

[77] Michael L Seltzer and Jasha Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6965–6969.

[78] Dongpeng Chen, Brian Mak, Cheung-Chi Leung, and Sunil Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5592–5596.

[79] Andrew R Barron, "Approximation and estimation bounds for artificial neural networks," *Journal of Machine learning*, vol. 14, no. 1, pp. 115–133, 1994.

[80] Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab, "Using resources from a closely-related language to develop ASR for

a very under-resourced language: A case study for Iban," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2015.

[81] Andreas Stolcke, "SRILM-an extensible language modeling toolkit," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2002, pp. 901–904.

[82] Paul Boersma and D Weenik, "PRAAT: a system for doing phonetics by computer. Report of the Institute of Phonetic Sciences of the University of Amsterdam," *Amsterdam: University of Amsterdam*, 1996.

[83] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[84] Dean Luo, Xuesong Yang, and Lan Wang, "Improvement of segmental mispronunciation detection with prior knowledge extracted from large L2 speech corpus," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2011, pp. 1593–1596.

[85] Alissa M Harrison, Wing Yiu Lau, Helen M Meng, and Lan Wang, "Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2008, pp. 2787–2790.

[86] Wai Kit Lo, Shuang Zhang, and Helen M Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2010, pp. 765–768.

[87] Xiaojun Qian, Helen M Meng, and Frank K Soong, "On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (CAPT)," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2011, pp. 865–868.

[88] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, no. 2, pp. 83–93, 2000.

[89] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 95–108, 2000.

[90] Sandra Kanters, Catia Cucchiarini, and Helmer Strik, "The goodness of pronunciation algorithm: a detailed performance study," *International Workshop on Speech and Language Technology in Education (SLaTE)*, pp. 49–52, 2009.

[91] Helmer Strik, Khiet Truong, Febe De Wet, and Catia Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[92] LaDeana F Weigelt, Steven J Sadoff, and James D Miller, "Plosive/fricative distinction: the voiceless case," *The Journal of the Acoustical Society of America (JASA)*, vol. 87, no. 6, pp. 2729–2737, 1990.

[93] Paul Iverson and Bronwen G Evans, "Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers," *The Journal of the Acoustical Society of America (JASA)*, vol. 126, no. 2, pp. 866–877, 2009.

[94] Miwako Hisagi, Kanae Nishi, and Winifred Strange, "Acoustic properties of Japanese and English vowels: Effects of phonetic and prosodic context," *Japanese/Korean Linguistics*, vol. 13, pp. 223–224, 2008.

[95] Yingming Gao, Richeng Duan, Jinsong Zhang, and Yanlu Xie, "SVM-based mispronunciation detection of Chinese aspirated consonants (/p/, /t/, /k/) by Japanese native speakers," in *Proceedings of the 9th Phonetic Conference of China (PCC)*, 2014, pp. 193–196.

[96] Joost Van Doremalen, Catia Cucchiarini, and Helmer Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *Proceedings of the 11th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2009, pp. 580–585.

[97] Shen Huang, Hongyan Li, Shijin Wang, Jiaen Liang, and Bo Xu, "Automatic reference independent evaluation of prosody quality using multiple knowledge fusions," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2010, pp. 610–613.

[98] Frederik Stouten and Jean-Pierre Martens, "On the use of phonological features for pronunciation scoring," in *Proceedings of the 31st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2006, vol. I, pp. 329–332.

[99] Yow-Bang Wang and Lin-Shan Lee, "Toward unsupervised discovery of pronunciation error patterns using universal phoneme posteriorgram for computer-assisted language learning," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8232–8236.

[100] Christian Hacker, Tobias Cincarek, Andreas Maier, Andre Hebler, and Elmar Noth, "Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children," in *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2007, vol. IV, pp. 197–200.

[101] Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based SVM," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2009, pp. 1903–1906.

[102] Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat, "Landmark-based automated pronunciation error detection," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2010, pp. 614–617.

[103] Jialu Zhang, "Distinctive feature system in Mandarin Chinese," *Chinese Journal of Acoustics*, vol. 30, no. 6, pp. 506–514, 2006.

[104] Jialu Zhang, "Distinctive feature tree in Mandarin Chinese," *Chinese Journal of Acoustics*, vol. 31, no. 3, pp. 193–198, 2006.

[105] Mengjie Wang and Zihou Meng, "Classification of Chinese word-finals based on distinctive feature detection," *Proceedings of the 3rd International Symposium on ElectroAcoustic Technologies (ISEAT)*, vol. 35, no. 9, pp. 38–41, 2011.

[106] Wen Cao, Dongning Wang, Jinsong Zhang, and Ziyu Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2010, pp. 1922–1925.

[107] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al., "The HTK book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[108] Lan Wang, Xin Feng, and Helen M Meng, "Automatic generation and pruning of phonetic mispronunciations to support computer-aided pronunciation training," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2008, pp. 1729–1732.

[109] Yingming Gao, Yanlu Xie, Wen Cao, and Jinsong Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in *Proceedings of the 16th Annual Conference of the International*

Speech Communication Association (Interspeech). ISCA, 2015, pp. 693–696.

[110] Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, and Jinsong Zhang, "Landmark of Mandarin nasal codas and its application in pronunciation error detection," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5370–5374.

[111] Silke M Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. IS ADEPT*, vol. 6, 2012.

[112] Richeng Duan, Jinsong Zhang, Wen Cao, and Yanlu Xie, "A preliminary study on ASR-based detection of Chinese mispronunciation by Japanese learners," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2014, pp. 1478–1481.

[113] Nancy F Chen and Haizhou Li, "Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*. IEEE, 2016, pp. 1–7.

[114] Yingming Gao, Richeng Duan, Jinsong Zhang, and Yanlu Xie, "SVM-based mispronunciation detection of Chinese aspirated consonants (/p/,/t/,/k/) by Japanese native speakers," in *Proceedings of 9th Phonetic Conference of China (PCC)*, 2014, pp. 193–196.

[115] Mark Hasegawa-Johnson, James Baker, Steven Greenberg, Katrin Kirchhoff, Jennifer Muller, Kemel Sonmez, Sarah Borys, Ken Chen, Amit Juneja, Emily Coogan, Karen Livescu, Srividya Mohan, and Tianyu Wang, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Technical report of the Johns Hopkins Center for Language and Speech Processing*, 2005.

[116] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning (ICML)*. ACM, 2006, pp. 369–376.

[117] Jie Li, Heng Zhang, Xinyuan Cai, and Bo Xu, "Towards end-to-end speech recognition for Chinese Mandarin using long short-term memory recurrent neural networks," in *Sixteenth annual conference of the international speech communication association*, 2015.

[118] Yimeng Zhuang, Xuankai Chang, Yanmin Qian, and Kai Yu, "Unrestricted vocabulary keyword spotting using LSTM-CTC," in *Proceedings of the 17th*

Annual Conference of the International Speech Communication Association (Interspeech). ISCA, 2016, pp. 938–942.

[119] Girish Palshikar, "Simple algorithms for peak detection in time-series," in *Proceedings of the First International Conference on Advanced Data Analysis, Business Analytics and Intelligence*, 2009, vol. 122.

[120] George A Miller and Patricia E Nicely, "An analysis of perceptual confusions among some English consonants," *The Journal of the Acoustical Society of America (JASA)*, vol. 27, no. 2, pp. 338–352, 1955.

[121] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[122] Yajie Miao, Mohammad Gowayyed, and Florian Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proceedings of the 14th IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[123] Sri Harsha Dumpala, Bhanu Teja Nellore, Raghu Ram Nevali, Suryakanth V Gangashetty, and B Yegnanarayana, "Robust vowel landmark detection using epoch-based features," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2016, pp. 160–164.

[124] Sheng Gao, Bo Xu, Hong Zhang, Bing Zhao, Chengrong Li, and Taiyi Huang, "Update progress of Sinohear: advanced Mandarin LVCSR system at NLPR," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 2000, vol. 3, pp. 798–801.

[125] Kenneth N Stevens, "Modelling affricate consonants," *Speech Communication*, vol. 13, no. 1-2, pp. 33–43, 1993.

[126] Christine Helen Shadle, "The acoustics of fricative consonants," *The Journal of the Acoustical Society of America (JASA)*, vol. 79, no. 2, pp. 574–574, 1986.

[127] Fredericka Bell-Berti, "Control of pharyngeal cavity size for English voiced and voiceless stops," *The Journal of the Acoustical Society of America (JASA)*, vol. 57, no. 2, pp. 456–461, 1975.

[128] Court S Crowther and Virginia Mann, "Native language factors affecting use of vocalic cues to final consonant voicing in English," *The Journal of the Acoustical Society of America (JASA)*, vol. 92, no. 2, pp. 711–722, 1992.

[129] Pascal Auzou, Canan Ozsancak, Richard J Morris, Mary Jan, Francis Eustache, and Didier Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical linguistics & phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[130] Martin Cooke and Odette Scharenborg, "The interspeech 2008 consonant challenge," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2008, pp. 1765–1768.

[131] Nabil N Bitar and Carol Y Espy Wilson, "Speech parameterization based on phonetic features: application to speech recognition," in *Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech)*, 1995, pp. 1411–1414.

[132] AM Abdelatty Ali, Jan Van der Spiegel, and Paul Mueller, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," in *Proceedings of the 23rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1998, vol. II, pp. 961–964.

[133] Ahmed M Abdelatty Ali, Jan Van der Spiegel, and Paul Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 833–841, 2001.

[134] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[135] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[136] David Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[137] Kenneth N Stevens and Dennis H Klatt, "Role of formant transitions in the voiced-voiceless distinction for stops," *The Journal of the Acoustical Society of America (JASA)*, vol. 55, no. 3, pp. 653–659, 1974.

[138] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[139] Özgül Salor, Bryan Pellom, Tolga Ciloglu, Kadri Hacioglu, and Mübeccel Demirekler, "On developing new text and audio corpora and speech recognition tools for the Turkish language," in *Proceedings of the 7th International*

Conference on Spoken Language Processing (ICSLP). ISCA, 2002, pp. 349–352.

[140] Janet Holmes, *An introduction to sociolinguistics*, Routledge, 2013.

[141] Victor Soto, Olivier Siohan, Mohamed Elfeky, and Pedro Moreno, "Selection and combination of hypotheses for dialectal speech recognition," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5845–5849.

[142] Fadi Biadsy and Julia Hirschberg, "Using prosody and phonotactics in Arabic dialect identification," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2009, pp. 208–211.

[143] Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, and Austin Waters, "Towards acoustic model unification across dialects," in *Proceedings of the 6th IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 624–628.

[144] Kanishka Rao and Haşim Sak, "Multi-accent speech recognition with hierarchical grapheme based models," in *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4815–4819.

[145] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the KLD-regularized model adaptation," in *Proceedings of 15th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2014, pp. 2977–2981.

[146] Mingming Chen, Zhanlei Yang, Jizhong Liang, Yanpeng Li, and Wenju Liu, "Improving deep neural networks based multi-accent Mandarin speech recognition using i-vectors and accent-specific top layer," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2015, pp. 3620–3624.

[147] Jiangyan Yi, Hao Ni, Zhengqi Wen, Bin Liu, and Jianhua Tao, "CTC regularized model adaptation for improving LSTM RNN based multi-accent Mandarin speech recognition," in *Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium on*. IEEE, 2016, pp. 1–5.

[148] Yanli Zheng, Richard Sproat, Liang Gu, Izhak Shafran, Haolang Zhou, Yi Su, Daniel Jurafsky, Rebecca Starr, and Su-Youn Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," in *Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2005, pp. 217–220.

[149] Mohamed Elfeky, Pedro Moreno, and Victor Soto, "Multi-dialectical languages effect on speech recognition: Too much choice can hurt," in *International Conference on Natural Language and Speech Processing (ICNLSP)*, 2015.

[150] Janet Pierrehumbert, "Phonological representation: Beyond abstract versus episodic," *Annu. Rev. Linguist.*, vol. 2, pp. 33–52, 2016.

[151] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*. ACM, 2014, pp. 1764–1772.

[152] Haşim Sak, Andrew Senior, Kanishka Rao, and Françoise Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2015.

[153] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[154] Andrew Maas, Ziang Xie, Dan Jurafsky, and Andrew Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 345–354.

[155] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[156] Liang Lu, Xingxing Zhang, Kyunghyun Cho, and Steve Renals, "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2015, pp. 3249–3253.

[157] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[158] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of the 41st IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP),. IEEE, 2016, pp. 4960–4964.

[159] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, "Direct acoustics-to-word models for English conversational speech recognition," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017, pp. 959–963.

[160] Hagen Soltau, George Saon, and Brian Kingsbury, "The IBM Attila speech recognition toolkit," in *Proceedings of the 3rd IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2010, pp. 97–102.

[161] Nelson Morgan and Hervé Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," in *Advances in Neural Information Processing Systems (NIPS)*, 1990, pp. 630–637.

[162] *Frustratingly easy domain adaptation*, 2009.

[163] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Trans. Audio Speech and Language*, vol. 25, no. 12, pp. 2410–2423, 2017.

[164] Yajie Miao, Mohammad Gowayyed, Xingyu Na, Tom Ko, Florian Metze, and Alexander Waibel, "An empirical exploration of CTC acoustic models," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2623–2627.

[165] Haşim Sak, Andrew Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4280–4284.

[166] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[167] Kenneth N Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," *The Journal of the Acoustical Society of America (JASA)*, vol. 69, no. S1, pp. S116–S116, 1981.

[168] John J McCarthy, "Feature geometry and dependency: A review," *Phonetica*, vol. 45, no. 2-4, pp. 84–108, 1988.

[169] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceeding of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 249–256.

[170] Herve A Bourlard and Nelson Morgan, *Connectionist speech recognition: a hybrid approach*, vol. 247, Springer Science & Business Media, 2012.